Genetic and genomic investigations of amyotrophic lateral sclerosis

By

Emily Pamela McCann

A thesis submitted to Macquarie University
for the degree of
Doctor of Philosophy
Department of Biomedical Sciences
Faculty of Medicine and Health Sciences
April 2019

MACQUARIE
University

Examiner's Copy

Typeset in LaTeX 2$_\varepsilon$.

This thesis is submitted to Macquarie University in fulfilment of the requirement for the Degree of Doctor of Philosophy.

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Emily Pamela McCann

"Hold on tight, this ride is a wild one"

All Time Low - Missing you

Acknowledgements

First and foremost, my warmest heartfelt thanks goes to the ALS patients and their families who have selflessly volunteered to participate in this research. Your strength and determination are an inspiration, and our research would not be possible without you. I would also like to thank the Motor Neuron Disease Research Institute of Australia for their ongoing support of our laboratory's genetic research endeavours, and particularly for the PhD top-up scholarship supporting this project.

To my supervisor Ian, thank you for your constant support, guidance and patience. Your passion for solving ALS and all things genetics is truly inspiring. Thank you for allowing me the the opportunity to work on this amazing project and trusting me to do it justice. Your generosity in sharing your time and knowledge has been invaluable, and I am especially grateful for your ability to always rationalise the situation and give me the illusion that I had everything under control even when I certainly did not.

As for the rest of the Blair group, past and present, we have grown so much over the last few years, and I could write a whole other thesis about how I grateful I am to each and everyone of you. Jenn, it's hard to put into words how very thankful I am to have you as a mentor and friend. You have taught me so much about the wonderful world of research, and have always been there to lend a hand or an ear whenever I needed it. You have become an incredible post-doc, and I hope one day I can follow in your footsteps and do the same. Kelly, you are an amazing scientist and learning from you is a privilege. Thank you for sharing your knowledge and your ongoing support since I embarked on this research journey 5 odd years ago.

To the amazing post-docs Alison and Shu, you are both so talented and so modest, thank you for opening my eyes beyond genetics to the wonderful worlds of zebrafish and cells! As for our fabulous research assistants Natalie G, Ingrid, Katharine, Sarah Frecks and Jasmin, you guys are the powerhouse of our lab and nothing would get

done without you! Special mention for those of you who helped out with my (what felt like) millions of PCRs! Huge thank you also goes to my student buddies Sandrine, Owen and Sharlynn for keeping me company in the dark corner of the office. Thank you for our (soul cleansing) conversations about all things random, and my deepest apologies for going off on tangents and probably never getting to the point when you asked me questions. To the amazing Biobank duo, Sarah and Elisa, our research would never get done without tour tireless efforts, thank you for making this all possible! To our bioinformatician Natalie T, thank you for your help with scripting things and getting me on the cluster so I could actually get my work done! On a group-wide note, thank you for all the teas and delicious baked goods that make it worth coming into the office for!

On a more personal note, I must express my appreciation for my amazing family. Mum and Dad, thank you for supporting my never ending journey of learning. I would not be the person I am today if it was not for you, and I am so very lucky that you have always encouraged me to follow my heart and reach for the stars. To my brother Shaun, if it wasn't for you I never would have believed I could learn how to code. Thank you for always making me see the lighter side of life and your occasional technical assistance. Callum, you have been a shoulder to cry on, my biggest cheer leader and the voice of reason too many times to count throughout this journey. Thank you for believing in me when I didn't believe in myself.

List of Publications

Publications/manuscripts presented as part of this thesis

Paper I: **McCann EP**, Williams KL, Fifita JA, Tarr I, O'Connor J, Rowe DB, Nicholson GA, Blair IP. *The genotype-phenotype landscape of familial amyotrophic lateral sclerosis in Australia.* Clin. Genet. **92**(3):259-266. (2017)

Manuscript II: **McCann EP***, Fifita JA*, Galper J, Grima N, Hogan AL, Freckleton S, Zhang KY, Chan Moi Fat S, Mehta P, Jagaraj CJ, Williams KL, Twine NA, Bauer DC, Kowk J, Halliday G, Walker AK, Atkin J, Rowe DB, Nicholson GA, Yang S∧, Blair IP∧. *Genetic and immunopathological analysis of CHCHD10 in Australian amyotrophic lateral sclerosis and frontotemporal dementia.* Prepared for submission to J Neurol Neurosurg Psychiatry.
* denotes equal first authorship.
∧ denotes equal senior authorship.

Manuscript III: **McCann EP**, Twine NA, Bauer DC, Grima N, Nicholson GA, Rowe DB, Blair IP, Williams KL. *Whole genome sequencing of amyotrophic lateral sclerosis discordant monozygotic twins identifies thousands of false positive de novo mutations.* Prepared for submission to Hum. Genomics.

Additional publications/manuscripts presented in this thesis

Paper A1: van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, van der Spek RA, Vsa U, de Jong S, Robinson MR, Yang J, Fogh I, van Doormaal PT, Tazelaar GH, Koppers M, Blokhuis AM, Sproviero W, Jones AR, Kenna KP, van Eijk KR, Harschnitz O, Schellevis RD, Brands WJ, Medic J, Menelaou A, Vajda A, Ticozzi N, Lin K, Rogelj B, Vrabec K, Ravnik-Glava M, Koritnik B, Zidar J, Leonardis L, Groelj LD, Millecamps S, Salachas F, Meininger V, de Carvalho M, Pinto S, Mora JS, Rojas-Garca R, Polak M, Chandran S, Colville S, Swingler R, Morrison KE, Shaw PJ, Hardy J, Orrell RW, Pittman A, Sidle K, Fratta P, Malaspina A, Topp S, Petri S, Abdulla S, Drepper C, Sendtner M, Meyer T, Ophoff RA, Staats KA, Wiedau-Pazos M, Lomen-Hoerth C, Van Deerlin VM, Trojanowski JQ, Elman L, McCluskey L, Basak AN, Tunca C, Hamzeiy H, Parman Y, Meitinger T, Lichtner P, Radivojkov-Blagojevic M, Andres CR, Maurel C, Bensimon G, Landwehrmeyer B, Brice A, Payan CA, Saker-Delye S, Drr A, Wood NW, Tittmann L, Lieb W, Franke A, Rietschel M, Cichon S, Nthen MM, Amouyel P, Tzourio C, Dartigues JF, Uitterlinden AG, Rivadeneira F, Estrada K, Hofman A, Curtis C, Blauw HM, van der Kooi AJ, de Visser M, Goris A, Weber M, Shaw CE, Smith BN, Pansarasa O, Cereda C, Del Bo R, Comi GP, D'Alfonso S, Bertolin C, Sorar G, Mazzini L, Pensato V, Gellera C, Tiloca C, Ratti A, Calvo A, Moglia C, Brunetti M, Arcuti S, Capozzo R, Zecca C, Lunetta C, Penco S, Riva N, Padovani A, Filosto M, Muller B, Stuit RJ; PARALS Registry; SLALOM Group; SLAP Registry; FALS Sequencing Consortium; SLAGEN Consortium; NNIPPS Study Group, Blair I, Zhang K, **McCann EP**, Fifita JA, Nicholson GA, Rowe DB, Pamphlett R, Kiernan MC, Grosskreutz J, Witte OW, Ringer T, Prell T, Stubendorff B, Kurth I, Hbner CA, Leigh PN, Casale F, Chio A, Beghi E, Pupillo E, Tortelli R, Logroscino G, Powell J, Ludolph AC, Weishaupt JH, Robberecht W, Van Damme P, Franke L, Pers TH, Brown RH, Glass JD, Landers JE, Hardiman O, Andersen PM, Corcia P, Vourc'h P, Silani V, Wray NR, Visscher PM, de Bakker PI, van Es MA, Pasterkamp RJ, Lewis CM, Breen G, Al-Chalabi A, van den Berg LH, Veldink JH. *Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis.* Nat. Genet. **48**(9):1043-8. (2016)

Paper A2: Benyamin B, He J, Zhao Q, Gratten J, Garton F, Leo PJ, Liu Z, Mangelsdorf M,

Al-Chalabi A, Anderson L, Butler TJ, Chen L, Chen XD, Cremin K, Deng HW, Devine M, Edson J, Fifita JA, Furlong S, Han YY, Harris J, Henders AK, Jeffree RL, Jin ZB, Li Z, Li T, Li M, Lin Y, Liu X, Marshall M, **McCann EP**, Mowry BJ, Ngo ST, Pamphlett R, Ran S, Reutens DC, Rowe DB, Sachdev P, Shah S, Song S, Tan LJ, Tang L, van den Berg LH, van Rheenen W, Veldink JH, Wallace RH, Wheeler L, Williams KL, Wu J, Wu X, Yang J, Yue W, Zhang ZH, Zhang D, Noakes PG, Blair IP, Henderson RD, McCombe PA, Visscher PM, Xu H, Bartlett PF, Brown MA, Wray NR, Fan D. *Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus with amyotrophic lateral sclerosis.* Nat Commun. **8**(1):611. (2017)

Paper A3: Fifita JA, Zhang KY, Galper J, Williams KL, **McCann EP**, Hogan A, Saunders N, Bauer D, Tarr IS, Pamphlett R, Nicholson GA, Rowe D, Yang S, Blair IP. *Genetic and Pathological Assessment of hnRNPA1, hnRNPA2/B1, and hnRNPA3 in Familial and Sporadic Amyotrophic Lateral Sclerosis.* NNeurodegener Dis. **17**(6):304-312. (2017)

Manuscript A4: Tarr IS, **McCann EP**, Benyamin B, Peters TJ, Twine NA, Zhang KY, Zhao Q, Zhang ZH, Rowe DB, Nicholson GA, Bauer DC, Clark SJ, Blair IP and Williams KL. *Monozygotic twins and triplets discordant for amyotrophic lateral sclerosis display differential methylation and gene expression.* Prepared for submission to Clin. Epigenetics.

Additional publications during candidature

Thomas-Jinu S, Gordon PM, Fielding T, Taylor R, Smith BN, Snowden V, Blanc E, Vance C, Topp S, Wong CH, Bielen H, Williams KL, **McCann EP**, Nicholson GA, Pan-Vazquez A, Fox AH, Bond CS, Talbot WS, Blair IP, Shaw CE, Houart C. *Non-nuclear Pool of Splicing Factor SFPQ Regulates Axonal Transcripts Required for Normal Motor Development.* Neuron. **94**(2):322-336. (2017)

Fifita JA, Williams KL, Sundaramoorthy V, **McCann EP**, Nicholson GA, Atkin JP, Blair IP. *A novel amyotrophic lateral sclerosis mutation in OPTN induces ER stress and Golgi fragmentation in vitro.* Amyotroph. Lateral Scler. Frontotemporal Degener. **18**(1-2):126-133. (2017)

Kichkin E, Visvanathan A, Lovicu FJ, Shu DY, Das SJ, Reddel SW, **McCann EP**, Zhang KY, Williams KL, Blair IP, Phillips WD *Postnatal Development of Spasticity Following Transgene Insertion in the Mouse IV Spectrin Gene (SPTBN4)* . J Neuromuscul Dis. **4**(2):159-164. (2017)

Williams KL, Topp S, Yang S, Smith B, Fifita JA, Warraich ST, Zhang KY, Farrawell N, Vance C, Hu X, Chesi A, Leblond CS, Lee A, Rayner SL, Sundaramoorthy V, Dobson-Stone C, Molloy MP, van Blitterswijk M, Dickson DW, Petersen RC, Graff-Radford NR, Boeve BF, Murray ME, Pottier C, Don E, Winnick C, **McCann EP**, Hogan A, Daoud H, Levert A, Dion PA, Mitsui J, Ishiura H, Takahashi Y, Goto J, Kost J, Gellera C, Gkazi AS, Miller J, Stockton J, Brooks WS, Boundy K, Polak M, Muoz-Blanco JL, Esteban-Prez J, Rbano A, Hardiman O, Morrison KE, Ticozzi N, Silani V, de Belleroche J, Glass JD, Kwok JB, Guillemin GJ, Chung RS, Tsuji S, Brown RH Jr, Garca-Redondo A, Rademakers R, Landers JE, Gitler AD, Rouleau GA,, Cole NJ, Yerbury JJ, Atkin JD, Shaw CE, Nicholson GA, Blair IP. *CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia.* Nat. Communications **7**:11253. (2016)

Williams KL, **McCann EP**, Fifita JA, Zhang K, Duncan EL, Leo PJ, Marshall M, Rowe DB, Nicholson GA, Blair IP.*Novel TBK1 truncating mutation in a familial amyotrophic lateral sclerosis patient of Chinese origin.* Neurobiol. Aging **36**(12):3334.e1-3334.e5. (2015)

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal, genetically heterogeneous neurodegenerative disease characterised by the loss of upper and lower motor neurons. Gene mutations remain the only proven cause of ALS. While 10% of patients have a family history (familial ALS; FALS), one third of these patients carry an unidentified causal mutation. Among the remaining 90% of apparently sporadic patients (sporadic ALS; SALS), less than 10% carry a known causal mutation. As such, a significant amount of genetic variation underlying ALS remains to be discovered.

This thesis presents innovative approaches to identify novel genetic causes of ALS using next-generation sequencing (NGS). This involved the development and application of various bioinformatics strategies to whole-exome (WES) and whole-genome (WGS) sequencing datasets for various patient cohorts including FALS patients, families and ALS-discordant monozygotic twins. Assessment of the prevalence of known and candidate ALS genes among Australian patients revealed that 39.2% of FALS had an unidentified causal gene mutation, and identified eight candidate ALS mutations. Novel ALS gene discovery in four small families identified 19, 11, 16 and 64 candidate causal mutations in each. Having exhausted the genetic power of these families, an *in silico* pipeline was developed to assess the potential pathogenicity of each candidate mutation. This showed that five, six, one and 11 candidate mutations had a high potential to cause ALS. Gene discovery efforts in a fifth family using WES, WGS and genetic linkage data failed to identify any candidate mutations, however narrowed the search to just 14% of the genome. WGS of four ALS-discordant monozygotic twin sets also failed to identify any *de novo* mutations underlying disease discordance. This work expands our understanding of the genetic causes of ALS, and in turn provides much needed insight for the development of diagnostic and carrier-screening regimes, as well as relevant models of disease.

Abbreviations

5mhC	5-Hydroxymethylcytosine
AC	Alternate allele count
ACMG	American College of Medical Genetics and Genomics
AD	Autosomal dominant inheritance
ALS	Amyotrophic lateral sclerosis
ALSdb	ALS Data Browser
ALT	Alternate allele
AN	Total allele count
AR	Autosomal recessive inheritance
AVS	ALS Variant Server
BAM	Binary Alignment/Map
BioGrid	Biological General Repository for Interaction Datasets
BMAA	Beta-Methylamino-L-Alanine
bp	Base pair
BWA	Burrows Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
Chr	Chromosome
CHROM	Chromosome

cM	Centimorgans
CNV	Copy number variant (Variation)
CSIRO	Commonwealth Scientific and Industrial Research Organisation
dbGAP	Database of Genotypes and Phenotypes
dbNSFP	Database for Non-Synonymous Snps' Functional Predictions
DENN	Differentially expressed in normal and neoplasia
DNA	Deoxyribonucleic acid
DPR	Dipeptide repeat (proteins)
DZ	Dizygotic
ER	Endoplasmic reticulum
ExAC	Exome Aggregation Consortium
FALS	Familial amyotrophic lateral sclerosis
FTD	Frontotemporal dementia
GATK	Genome Analysis ToolKit
gDNA	Genomic DNA
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GQ	Genotype quality
GTex	Genotype-Tissue Expression Project
GWAS	Genome-wide association study (studies)
HBT	Human Brain Transcriptome
HPCC	High performance computing cluster
ID	Variant identity
IF	Immunofluorescence

IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
indel	Insertion/deletion
INFO	Annotation information
kb	Kilobase
LMN	Lower motor neuron
LOD	Logarithm of odds
LOH	Loss of heterozygosity
MAF	Minor allele frequency
Mb	Megabases
MGRB	Medical Genome Reference Bank
miRNA	Micro RNA
MND	Motor neuron disease
mRNA	Messenger RNA
MZ	Monozygotic
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding RNA
NFE	Non-Finnish European
NGS	Next-generation sequencing
NHLBI-ESP	National Heart, Lung, and Blood Institute - Exome Sequencing Project
PBP	Progressive Bulbar Palsy
PCR	Polymerase chain reaction
PhastCons	Phylogenetic Analysis With Space/Time Models - Conservation

PhyloP	Phylogenetic Model
PLS	Primary lateral sclerosis
PMA	Progressive muscular atrophy
Polyphen-2	Polymorphism Phenotyping V2
Pon-P2	Pathogenic-or-not-pipeline
POS	Genomic DNA position
PROVEAN	Protein Variation Effect Analyzer
pVAAST	Pedigree Variant Annotation, Analysis and Search Tool
PXX	Proline-X-X aminio acid sequence
QUAL	Variant quality
RAN	Repeat-associated non-AUG (translation)
REF	Reference allele
RGG1-3	Arginine Glycine Glycine repeat region
RNA	Ribonucleic acid
RPKM	Reads per kilobase of transcript per million
RRM	RNA-recognition motif
RVIS	Residual variation intolerance
SALS	Sporadic amyotrophic lateral sclerosis
SAM	Sequence Alignment/Map
SEA	South East Asian
SIFT	Sorting Intolerant From Tolerant
SMA	Spinal muscular atrophy
SMART	Simple Modular Architecture Research Tool
SNP	Single nucleotide polymorphism

SNV	Single nucleotide variant
SV	Strucural variant (variation)
SYGQ	Rich in Serine, Tyrosine, Glycine, Glutamine
UBA	Ubiquitin-associated domain
UBL	Ubiquitin-like domain
UCSC	University of California Santa Cruz
UMN	Upper motor neuron
UPS	Ubiquitin-proteasome system
UTR	Untranslated region
VCF	Variant Call File
WES	Whole-exome sequencing
WGA	Whole-genome amplified
WGS	Whole genome sequencing
XD	X-linked dominant inheritance

Contents

List of Figures

List of Tables

"Quiet people have the loudest minds"

Stephen Hawking

1 Introduction

1.1 General introduction

Amyotrophic lateral sclerosis (ALS; also known as motor neuron disease, MND) is a fatal, late onset neurodegenerative disease caused by the death of the upper and lower motor neurons of the motor cortex, brain stem and spinal cord. Patients experience progressive muscle weakness, wasting and spasticity, eventually losing gross and fine motor capabilities to the point that they can no longer walk, speak, eat or breathe unassisted. Within just two to five years of symptom onset, most patients die from associated respiratory failure. There are no effective treatments for ALS. The only pharmaceutical approved in Australia for the treatment of ALS is riluzole, which only extends life by a matter of months. There is a drastic need for the development of more effective treatments for this devastating disease, which requires the identification of suitable lifestyle or drug targets. While many lifestyle factors and exposures have been suggested to cause or influence the onset of ALS, to-date, genetically inherited mutations remain the only proven cause of the disease. More than 20 genes have been shown to harbour mutations that cause ALS, and many more genes have been found to carry genetic variants associated with increased disease-risk. A small proportion of ALS patients have a family history of disease, while the remaining cases have seemingly sporadic onset. The majority of the known ALS causal mutations were discovered

by studying ALS families. These families were typically large, and amenable to classical genetic linkage analysis which facilitated disease gene identification. However, only two thirds of familial ALS patients carry a known ALS mutation, while this figure is a mere 10% for sporadic ALS patients. This leaves the cause of ALS in the majority of patients unsolved. This thesis presents strategies for the discovery of novel genetic causes of ALS in an era where the common genetic causes of disease have already been identified. Following the great success of ALS gene hunting in large ALS families, those ALS families remaining to have their causal mutations identified are genetically small. The genetic power of these small families for novel disease gene discovery is markedly decreased, as there is limited availability of DNA samples caused by the reduced penetrance of their causal mutations. This renders genetic linkage analysis in these families exceedingly difficult, and in some cases impossible. Therefore, large-scale whole exome and genome sequencing approaches are required to identify disease causal mutations in these families. As such, the remaining genetic causes of ALS lay hidden within such datasets, which are both large and complex, harbouring not only variants that contribute to the cause or predisposition to ALS, but also a plethora of benign variation masking the pathogenic culprits. In this thesis, pipelines have been developed to effectively handle the immense volume of genetic data generated by whole exome and genome sequencing as part of the search for the remaining genetic variation contributing to the cause of ALS. The following chapters detail the use of this genetic data for identifying novel gene variants that cause or are associated with ALS using candidate gene, family and twin based approaches, and the extension of these findings to sporadic patients. Each such ALS gene discovery will broaden the spectrum of known ALS genes, further our understanding of disease biology, and provide new targets for the development of cell and animal models, diagnostics and therapeutics.

1.2 What is MND?

Motor neuron disease (MND) is an umbrella term for a group of disorders characterised by the progressive degeneration and eventual death of motor neurons, that leads to various motor impairments in patients. Motor neuron death involves the degeneration of motor nerve axons and the destruction of neuromuscular junctions, causing a breakdown of communication with the muscle fibres innervated by these axons (Tiryaki and Horak, 2014).

Motor neurons are responsible for voluntary muscle movement. Those originating in the motor region of the cerebral cortex are known as the upper motor neurons (UMNs), and function by transmitting electrical impulses or "messages" to the lower motor neurons (LMNs) (Kiernan et al., 2011). LMNs originate in the brain stem (bulbar motor neurons) and innervate muscles involved in movements of the face and tongue, and control speaking, chewing and swallowing. Those LMNs that originate in the spinal cord (anterior horn cells) innervate larger limb muscles that control movements such as walking and writing (Kiernan et al., 2011). Figure 1.1 loosely depicts these characteristics.

Loss of UMNs leads to muscle spasticity, weakness and brisk deep tendon reflexes, while LMN loss is generally associated with muscle fasciculation, cramps, wasting and weakness (Kiernan et al., 2011; Swinnen and Robberecht, 2014; Tiryaki and Horak, 2014). The two restricted MND phenotypes, primary lateral sclerosis (PLS) and progressive muscular atrophy (PMA), involve either purely UMNs or LMNs, respectively. Disease progression rates vary drastically between these two restricted phenotypes, with some PLS patients living with slowly progressive disease for up to twenty years, while disease progresses more rapidly for PMA patients with typical survival at just five years. Progressive bulbar palsy (PBP) occurs when the bulbar motor neurons are exclusively lost, which may involve either UMNs, LMNs or both (Al-Chalabi and Hardiman, 2013; Kiernan et al., 2011).

The most common of the MNDs is amyotrophic lateral sclerosis (ALS), which affects both the upper and lower motor neurons, with symptoms experienced in both the limbs and bulbar muscles (Al-Chalabi and Hardiman, 2013; Kiernan et al., 2011). Most ALS patients die within two to five years of first symptom onset, usually as a result of associated respiratory failure (Huisman et al., 2011). PLS, PMA and PBP can all progress to ALS, usually within the first few years after onset (Al-Chalabi and Hardiman, 2013). A visual summary of the various MND types is presented in Figure 1.1.

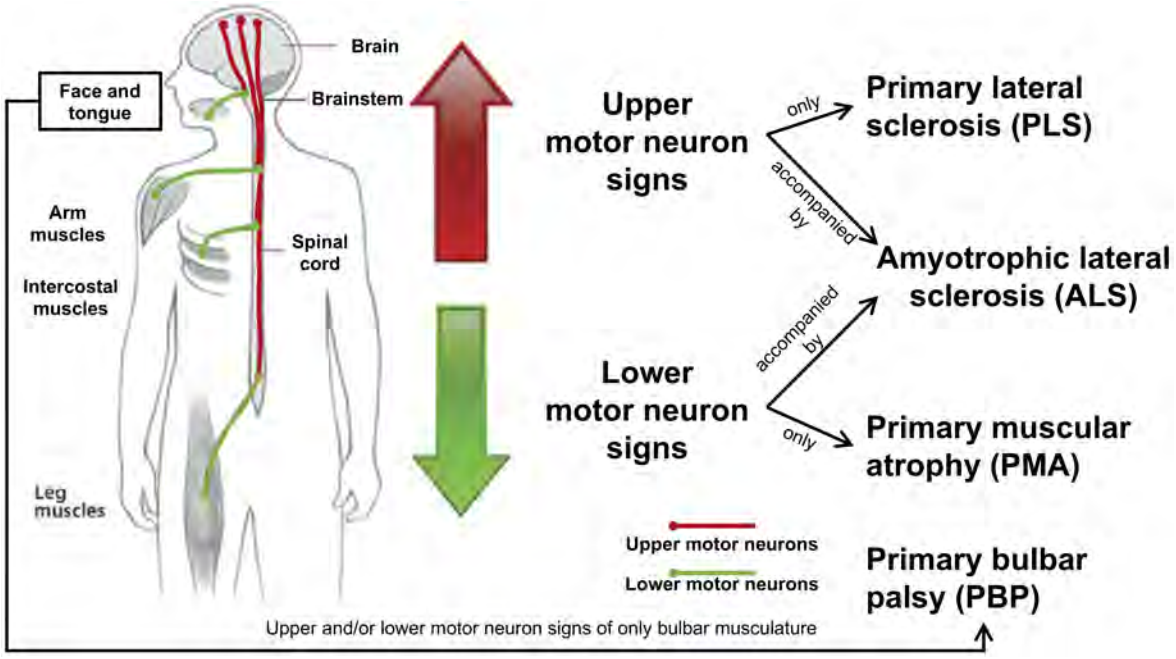


FIGURE 1.1: **Subtypes of motor neuron disease.** There are four subtypes of MND, defined by the involvement of either the upper or lower motor neurons, or both, and the affected musculature. PLS; primary lateral sclerosis, ALS; amyotrophic lateral sclerosis, PMA; progressive muscular atrophy and PBP; progressive bulbar palsy. Figure adapted from Tiryaki and Horak (2014).

1.3 Amyotrophic lateral sclerosis

1.3.1 Clinical features

Great variation is observed among ALS patients in terms of age and site of onset, rate of progression and prognosis (Ravits and La Spada, 2009). Onset of first symptoms can occur anywhere between the second and ninth decade of life, though is most often seen between the ages of 50 to 60 years (Swinnen and Robberecht, 2014). In rare cases, juvenile ALS is seen in patients under 25 years of age (Swinnen and Robberecht, 2014). Patients with a family history of ALS often have a younger age of onset, with a mean of 46 years, while sporadic patients have a mean age of onset at 56 years (Tiryaki and Horak, 2014). Most patients experience their first symptoms in the limbs (~70%), while others have bulbar onset (~25%) and in rare cases onset occurs in the trunk (Tiryaki and Horak, 2014). As described in Section 1.2, patients experience a range of muscular symptoms according to the type of motor neuron involvement (Kiernan et al., 2011; Swinnen and Robberecht, 2014; Tiryaki and Horak, 2014). Importantly, a key feature of disease is the progression and spread of these symptoms (Kiernan

et al., 2011; Swinnen and Robberecht, 2014). Disease progression is generally very rapid, with 50% of ALS patients dying within 30 months of onset, and a mere 20% surviving beyond five years. Shorter survival is associated with older age of onset, bulbar onset, as well as early onset of respiratory symptoms (Pupillo et al., 2014; Talbot, 2009). Conversely, predominately UMN involvement, younger age of onset and delayed formal diagnosis predict longer survival (Pupillo et al., 2014; Talbot, 2009).

Many ALS patients also suffer from some form of cognitive impairment, with the figure estimated to be as high as 50% (Montuschi et al., 2015; Ringholz et al., 2005). Reports show that 20-25% of ALS patients experience executive impairment, while 5-10% exhibit non-executive impairments such as language and memory deficits (Elamin et al., 2013; Montuschi et al., 2015). Most importantly, 10-15% of ALS patients meet the criteria for co-morbid frontotemporal dementia (FTD). These two conditions are considered to be a spectrum of neurodegenerative disease, owing to the significant co-morbidity between the two, as well as their shared genetic basis, and the similarities between the observed pathologies in affected neurons (Montuschi et al., 2015; Phukan et al., 2012; Ringholz et al., 2005).

1.3.2 Epidemiology

ALS is classed as a rare disease, with an estimated worldwide incidence of 1-2 individuals per 100,000 (Marin et al., 2017). However, this figure varies significantly between populations, and is far greater in Europe and North America compared with Asia (Marin et al., 2017). In Australia the estimated prevalence is 8.7 individuals per 100,000 (MND Australia; `www.mndaust.an.au`). The cumulative lifetime risk for ALS is approximately 1 in 300 (Johnston et al., 2006), and it is estimated that by 2040, there will be 400,000 ALS patients across the world (Blasco et al., 2016). Men are more commonly affected than women, with a male to female ratio of 1.6 to 1 (Tiryaki and Horak, 2014), however in familial cases, this ratio approaches 1 to 1 (Brown and Al-Chalabi, 2017). Approximately 10% of patients have a relative also affected by ALS (familial ALS; FALS), while the remaining 90% have no apparent family history of disease (sporadic ALS; SALS) (Brown and Al-Chalabi, 2017).

While there is still no clear consensus in the literature, it has been suggested that the incidence of ALS has increased in recent decades (Ingre et al., 2015), implicating environmental influences on the onset of ALS. Patient exposures to different elements,

chemicals or toxins, and participation in particular activities have been investigated as potentially predisposing individuals to developing ALS, however, to-date, none have been definitively shown to increase disease-risk.

A range of environmental factors such as pesticides, Beta-methylamino-L-alanine (BMAA), heavy metals, viruses, physical activity, body mass index, smoking and military service have been investigated in ALS patients. While many studies have suggested correlations between these factors and the incidence of disease, evidence against association has just as often been reported. This lack of consistency casts doubt over the link existing between these factors and disease (reviewed in Bozzoni et al., 2016; Ingre et al., 2015; Oskarsson et al., 2015; Trojsi et al., 2013). It is exceedingly difficult to determine which environmental factors are truly associated with disease and which associations are purely circumstantial (Brown and Al-Chalabi, 2017). This stems from there being a plethora of possible environmental risk agents, their potential to interact with each other and with genetic risk factors, as well as a probable biased representation of patients with a longer disease course presenting at clinics (Brown and Al-Chalabi, 2017). As such, future studies investigating environmental contributions to disease-risk need to be expanded to larger cohorts and include patients with a full range of ALS phenotypes. Further, as these environmental contributions are likely to interact with genetic ALS risk factors, they may partially explain the phenotypic variability observed between patients, particularly those carrying identical causal gene mutations.

1.3.3 Treatment

Despite over 60 molecules having been investigated as ALS drug treatments, there are still no pharmaceuticals available that markedly improve life expectancy or quality of life for ALS patients (Petrov et al., 2017). To-date, the most successful pharmaceutical intervention has been riluzole, an anti-glutamate agent that blocks glutamate related excitotoxicity through its inactivation of sodium channels (Bryson et al., 1996), though whether this property underlies its therapeutic action in ALS remains unknown. The first clinical trial of riluzole began in 1990 and reported marginal improvements to survival (Bensimon et al., 1994). However, it is widely accepted that the effect of riluzole is quite modest, slowing disease progression to extend survival by only two to three months (Miller et al., 2007). More recently, edaravone was approved as an ALS treatment in the USA. After originally being developed to treat stroke, edaravone was

trialled in ALS owing to its free-radical scavenging behaviour. The hypothesis was that the removal of free radicals may have a protective effect on motor neurons, in accordance with the role of the SOD1 protein in free-radical processing, and the major role of *SOD1* gene mutations in ALS (discussed in Section 1.4.1.1). While edaravone has been demonstrated to improve patient mobility, its effect on survival remains to be seen (Abe et al., 2017), and clinical trials suggest that beneficial effects may be limited to a small subset of patients meeting strict genetic criteria (Kiernan, 2018). Though we are yet to find a broadly applicable drug treatment for ALS, there are a number of promising clinical trials in progress, including antisense oligonucleotides for *SOD1* (Miller et al., 2013) and *C9orf72* (Donnelly et al., 2013; Riboldi et al., 2014).

As we wait for the development of more effective pharmaceutical interventions targeting the effectors of disease, ALS patients have the option to use medical equipment and other strategies to improve their comfort while living with disease. The best outcomes are seen when a multidisciplinary approach is taken for patient care. This involves a range of health professionals including specialist neurologists, nurses, physiotherapists, occupational therapists, neuropsychologists, speech therapists, respiratory physicians and gastroenterologists. By utilising these different disciplines, symptoms may be alleviated so that patients are able to experience a better quality of life than they would otherwise (Kiernan, 2018; Turner and Kiernan, 2015).

1.3.4 Pathology

Death of both the UMNs and LMNs is the defining pathological feature of ALS (Brown and Al-Chalabi, 2017). As the corticospinal neurons (those UMNs projecting from cortical regions through the brainstem into the spinal tract) degenerate, their descending axons in the lateral spinal cord become scarred (sclerosis), and as the spinal motor neurons die, secondary denervation occurs causing muscle wasting (amyotrophy) (Taylor et al., 2016).

The hallmark pathological feature of post-mortem ALS patient tissue is the presence of ubiquitinated protein aggregates, or inclusions, of misfolded proteins in the affected motor neurons (Forman et al., 2004; Leigh et al., 1989). In most patients, these protein aggregates contain the transactive response DNA-binding protein 43 (TDP-43). However, TDP-43 is absent from protein inclusions observed in patients with causal mutations in *SOD1* or *FUS* (discussed in Section 1.4.1). In patients with

TDP-43 pathology, the TDP-43 protein is typically cleaved, hyperphosphorylated, and mislocalised to the cytoplasm (Neumann et al., 2006). A number of other proteins have also been found within these neuronal inclusions, such as ubiquilin 2 (UBQLN2; Deng et al., 2011), fused in sarcoma (FUS; Neumann et al., 2011) and sequestosome 1 (p62; Taylor et al., 2016), among others. Some patients with TDP-43 pathology also have additional, unique pathological features. For example, patients carrying a pathogenic expansion in the *C9orf72* gene also have aggregates of dipeptide repeat proteins in the cerebellum and hippocampus (Brown and Al-Chalabi, 2017). Bunina bodies, or eosinophilic intraneuronal inclusions, in the remaining lower motor neurons are considered another hallmark feature of the disease (Okamoto et al., 2008).

1.3.5 Concepts of ALS pathogenesis

Both genetic and phenotypic heterogeneity are abundant among ALS patients, and the underlying pathological mechanisms of disease remain to be defined. However, there is a distinct convergence of the molecular processes implicated as playing an important role in ALS pathogenesis. These include RNA misprocessing, disrupted protein homeostasis, excitotoxicity, endoplasmic reticulum (ER) stress, Golgi fragmentation, and mitochondrial dysfunction (Brown and Al-Chalabi, 2017; Taylor et al., 2016; Therrien et al., 2016). The roles of these processes are not mutually exclusive, thus there is potential for the interplay between them to contribute to pathogenesis (Brown and Al-Chalabi, 2017). For example, abnormal RNA-binding proteins have been found within the protein aggregates seen in affected motor neurons, suggesting that degradation of RNA-binding proteins is compromised during the pathogenic process (Ling et al., 2013). Interestingly, many genes harbouring ALS causal mutations have roles in RNA processing and/or protein homeostasis (discussed in Sections 1.3.5.1 and 1.3.5.2).

1.3.5.1 RNA homeostasis and trafficking

The term RNA processing encompasses a range of different events including regulation of transcription and translation, pre-mRNA processing and splicing, and RNA transport. The key commonality underlying these processes is their reliance on RNA-binding proteins.

The major contribution of RNA processing defects to ALS pathogenesis was first

realised with the discovery of ALS mutations in two genes, *TARDBP* (encoding TDP-43) and *FUS*, both of which encode RNA-binding proteins. In addition to these seminal discoveries, other RNA-binding proteins including hnRNPs (Kim et al., 2013), TAF15 (Couthouis et al., 2011; Ticozzi et al., 2011), EWSR1 (Couthouis et al., 2012), ANG (Greenway et al., 2006), SETX (Chen et al., 2004) and ATXN2 (Elden et al., 2010) have also been implicated in ALS, reinforcing the crucial role of aberrant RNA processing in disease. These proteins have varied roles in gene splicing, microRNA (miRNA) production and axonal processes (Brown and Al-Chalabi, 2017). However, the commonalities between these proteins is not limited to their ability to bind RNA. They also share a propensity to bind other proteins through low complexity prion-like protein domains, in which most of their ALS causal mutations are located (Brown and Al-Chalabi, 2017; Kim et al., 2013; Robberecht and Philips, 2013). It appears that the ALS-linked mutations increase the binding propensity of these domains, causing self-binding and the formation of protein aggregates (Kim et al., 2013; Robberecht and Philips, 2013). Aggregates of RNA-binding proteins may also incorporate into stress granules, which are cytoplasmic complexes containing untranslated RNA transcripts encoding messenger ribonucleoproteins (Monahan et al., 2016; Protter and Parker, 2016).

The most common known cause of ALS, a hexanucleotide repeat expansion in *C9orf72* (see Section 1.4.1.4), has also been suggested to elicit its pathogenic effect through RNA-related processes. These hypotheses suggest the underlying pathogenic mechanism of expanded *C9orf72* alleles is mediated by either RNA foci or dipeptide repeat proteins (DPRs). RNA foci form when antisense expanded *C9orf72* transcripts are deposited within the nucleus, and subsequently sequester nuclear proteins (Zu et al., 2013). DPRs are highly prone to aggregation and are produced by repeat-associated non-AUG (RAN) translation of the C9orf72 repeat expansion (Ash et al., 2013; Mori et al., 2013a,b; Zu et al., 2013). A recent study by Kramer et al. (2018) investigating DPR toxicity found interplay between processes such as nucleo-cytoplasmic transport, RNA-processing pathways and chromatin modification, which affected the normal functioning of the ER and proteasome. This finding adds further support to the interaction between the different molecular processes implicated in ALS.

1.3.5.2 Protein homeostasis

Clearance of damaged, misfolded, aggregated and unnecessary proteins is imperative to maintain protein homeostasis for proper cellular function (Vilchez et al., 2014).

The two pathways that are primarily responsible for the degradation of abnormal proteins are the ubiquitin-proteasome system (UPS) and autophagy (Tanaka and Matsuda, 2014). Within the UPS, chaperone proteins recognise poly-ubiquitinated proteins that have been tagged for degradation and transport them to the proteasome to be unfolded and proteolysed (Finley, 2009). Autophagy is a normal physiological process dealing with the destruction of damaged proteins. However, it is induced and upregulated during periods of cellular stress, including ER stress, and in the presence of protein aggregates. It involves the formation of an autophagosome, which engulfs the damaged or dysfunctional protein, then fuses with a lysosome to form an autolysosome in which protein degradation occurs (Tanaka and Matsuda, 2014). During the human aging process, the efficiency of these systems decline and damaged proteins are more likely to accumulate (Vilchez et al., 2014).

As described in Section 1.3.4, protein aggregates are a key pathological feature of ALS. While it remains to be established whether protein aggregates in affected motor neurons are a cause of or consequence of disease, their presence implicates the important role of aberrant protein homeostasis in ALS (Therrien et al., 2016). Abnormal protein degradation is further implicated in disease by ALS-linked mutations in multiple genes encoding proteins that play key roles in protein degradation, including those involved with the UPS such as *UBQLN2* (Deng et al., 2011; Williams et al., 2012b), *OPTN* (Maruyama et al., 2010), *VCP* (Johnson et al., 2010) *TBK1* (Cirulli et al., 2015; Williams et al., 2015), *CCNF* (Williams et al., 2016b) and *SQSTM1* (Fecto et al., 2011), as well as autophagy related genes including *CHMP2B* (Parkinson et al., 2006; van Blitterswijk et al., 2012b) and *FIG4* (Chow et al., 2009).

1.4 Genetics of ALS

The genetic contribution to ALS was first acknowledged through the observation that a significant number of patients came from families where other members had also been diagnosed with ALS. It is generally accepted that approximately 10% of ALS patients have a family history of disease, and are classified as familial (Renton et al., 2014), although this figure can range from 5 to 20%, depending on the criteria used to define familial disease (Al-Chalabi et al., 2017, discussed in Section 1.6.1). The majority of ALS families show an autosomal dominant Mendelian pattern of inheritance, though reduced penetrance within families is often observed. The remaining 90% of ALS patients have no known family history of ALS. However, misclassification of ALS

patients is quite common, as ascertainment of a complete family history is often not possible beyond immediate family members. This is compounded by the fact that FALS and SALS are clinically and pathologically indistinguishable (Andersen and Al-Chalabi, 2011).

Genetic investigations of ALS patients and families have implicated over 50 genes in the disease, however the causality of many of these genes is questionable (Taylor et al., 2016). Table 1.1 summarises those genes that have the strongest evidence supporting their role in ALS pathogenesis. The genetic heterogeneity of ALS is evident with at least 25 genes harbouring mutations that cause ALS and a further 12 genes associated with disease. To-date, these genetic mutations are the only proven cause of ALS. Among Australian ALS, approximately 60% of FALS patients carry a known ALS mutation (Paper I, McCann et al., 2017), while this figure is approximately 10% for sporadic patients (Paper I, McCann et al., 2017, unpublished data). The genes harbouring mutations that cause ALS have provided the basis for most downstream ALS research, and have greatly enhanced our understanding of disease pathogenesis.

1.4.1 Familial ALS

Most ALS families carry autosomal dominant mutations that cause ALS, though X-linked and rare autosomal recessive mutations have also been reported. As a late onset disease with a highly variable age of onset, some family members who carry a causal ALS mutation die of other causes before they reach the age at which they may have developed disease. This means that incomplete penetrance is a common feature of many ALS family pedigrees.

The first gene to be identified with mutations that cause ALS was the copper/zinc superoxide dismutase gene, *SOD1* (Rosen, 1993). The next major breakthrough came 15 years later with the discovery of pathogenic mutations in *TARDBP*, which encodes the RNA-binding protein, TDP-43 (Sreedharan et al., 2008). TDP-43 had previously been identified as the principle component of the protein aggregates observed in post-mortem ALS patient motor neurons (Neumann et al., 2006). Soon after, mutations were identified in the *FUS* gene that encodes the fused in sarcoma protein, an RNA-binding protein functionally similar to TDP-43 (Kwiatkowski et al., 2009; Vance et al., 2009). In 2011, the most common known cause of ALS was identified as a hexanucleotide repeat expansion in the *C9orf72* gene (DeJesus-Hernandez et al., 2011; Renton et al., 2011). These four genes are the most common known ALS

Table 1.1: **Summary of currently known familial ALS genes.**

Gene symbol	Gene name	Gene locus	Inheritance	No. of mutations*	Estimated % FALS	Discovery	Pathway involvment	Reference
ALS2	Alsin	2q33.1	AR, juvenile	28	unknown	linkage	regeneration & motorneuronal death	(Hadano et al., 2001)
ANG	angiogenin	14q11.2	AD	29	<1%	candidate gene	altered DNA/RNA processing	(Greenway et al., 2006)
C9orf72	Chromosome 9 open reading frame 72	9p21.3p13.3	AD	1	40–50%	linkage, family NGS	altered DNA/RNA processing	(DeJesus-Hernandez et al., 2011; Renton et al., 2011)
CCNF	Cyclin F	16p13.3	AD	N/A	<1%	linkage, family NGS	cell cycle, protein ubiquitination	(Williams et al., 2016b)
CHCHD10	Coiled-coil-helix-coiled-coil-helix domain containing 10	22q11.23	AD	5	<1%	family NGS	mitochondria	(Bannwarth et al., 2014)
DCTN1	Dynactin 1	2p13.1	AD	7	N/A	candidate gene	axonal transport & vesicle trafficking	(Puls et al., 2003)
ERBB4	Erb-b2 receptor tyrosine kinase 4	2q34	AD	2	<1%	linkage, family NGS	neuregulin-ErbB4 pathway	(Takahashi et al., 2013)
FIG4	FIG4 phosphoinositide 5-phosphatase	6q21	AD	10	unknown	candidate gene	cell death	(Chow et al., 2009)
FUS	Fused in sarcoma	16p11.2	AD	80	1–5%	candidate gene	altered DNA/RNA processing	(Kwiatkowski et al., 2009; Vance et al., 2009)
GLE1	GLE1, RNA export mediator	9q34.11	AD	3	unknown	family NGS	altered DNA/RNA processing	(Kaneb et al., 2015)
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	12q13.13	AD	2	unknown	candidate gene, family NGS	altered DNA/RNA processing	(Kim et al., 2013)
MATR3	Matrin 3	5q31.2	AD	4	<1%	family NGS	altered DNA/RNA processing	(Johnson et al., 2014b)
OPTN	Optinuerin	10p15p14	AR, AD	39	<1%	homozygosity mapping	protein homeostasis	(Maruyama et al., 2010)
PFN1	Profilin 1	17p13.2	AD	12	unknown	family NGS	cytoskeleton & cellular transport	(Wu et al., 2012a)
SETX	Senataxin	9q34	AD, juvenile	8	<1%	linkage	altered DNA/RNA processing	(Chen et al., 2004)
SOD1	superoxide dismutase 1	21q22.1	AD #	185	20%	linkage	Oxidative stress	(Rosen, 1993)
SS18L1	Synovial sarcoma translocation gene on chr18-like 1	20q13.33	AD, *de novo*	3	unknown	family NGS	chromatin remodelling	(Chesi et al., 2013)
SQSTM1	Sequestosome 1	5q35	AD	17	N/A	candidate gene	proteostatic proteins	(Fecto et al., 2011)
TAF15	TATA-box binding protein associated factor 15	17q11.1q11.2	AD	7	unknown	candidate gene	altered DNA/RNA processing	(Couthouis et al., 2011)
TARDBP	TAR DNA binding protein	1p36.2	AD	53	1–5%	linkage, candidate gene	altered DNA/RNA processing	(Sreedharan et al., 2008)
TIA1	T-cell-restricted intracellular antigen-1	2p13.3	AD		<2%	family NGS	altered DNA/RNA processing	(Mackenzie et al., 2017)
TBK1	TANK-binding kinase 1	12q14.2	AD	18	<1%	NGS burden analysis	protein homeostasis	(Cirulli et al., 2015; Freischmidt et al., 2015)
UBQLN2	Ubiquilin 2	Xp11	XD	26	<1%	family NGS	protein homeostasis	(Deng et al., 2011)
VAPB	VAMP-associated protein B & C	20q13.3	AD	2	unknown	candidate gene	axonal transport & vesicle trafficking	(Nishimura et al., 2004)
VCP	Valosin containing protein	9p13	AD	7	<1%	candidate gene	protein homeostasis	(Johnson et al., 2010)
Genes associated with ALS								
ATXN2	Ataxin 2	12q24	AD	9	N/A	candidate gene	oxidative stress	(Elden et al., 2010)
C21orf2	Chromosome 21 open reading frame 2	21q22.3	sporadic	N/A	N/A	family NGS association	unknown	(van Rheenen et al., 2016)
CHMP2B	Chromatin modiftying protein 2B	3p11	AD	6	N/A	linkage, candidate gene	proteostatic proteins	(Parkinson et al., 2006)
DAO	D-amino-acid oxidase	12q24	AD	2	N/A	linkage, candidate gene	excitotoxicity	(Mitchell et al., 2010)
ELP3	Elongator acetyltransferase complex subunit 3	8p21.1	sporadic	0	N/A	mirosatellite, GWAS	projection neurons maturation	(Simpson et al., 2009)
GPX3-TNIP1	Glutathione peroxidase 3 & TNFAIP3 Interacting Protein 1	5q33.1	AD	N/A	N/A	GWAS	oxidative damage	(Benyamin et al., 2017)
NEFH	Neurofilament, heavy polypeptide	22q12.2	sporadic	11	N/A	candidate gene	cytoskeleton & microtubule	(Figlewicz et al., 1994)
NEK1	NIMA Related Kinase 1	4q33	-	N/A	N/A	NGS burden analysis	cell cycle regulation	(Kenna et al., 2016)
P4HB	prolyl 4-hydroxylase, beta polypeptide	17q25.3	-	N/A	N/A	association analysis	enzyme	(Kwok et al., 2013)
PRPH	Peripherin	12q13.12	sporadic	0	N/A	candidate gene	cytoskeleton	(Gros-Louis et al., 2004)
TUBA4A	Tubulin, alpha 4A	2q35	-	12	N/A	NGS burden analysis	cytoskeleton & microtubule	(Smith et al., 2014)
UNC13A	Protein Unc-13 Homolog A	19p13.11	sporadic	0	N/A	association analysis	regulates release of neurotransmitters	(van Es et al., 2009)

*The ALS Online Genetics Database (ALSoD), 2018 (http://alsod.iop.kcl.ac.uk/; Abel et al., 2013).
p.D90A mutation is recessive in Scandinavian populations.
Abbreviations: AD, autosomal dominant; AR, autosomal recessive; XD, X-linked dominant; N/A, absent; NGS, next-generation sequencing; and GWAS, genome-wide association study.

genes, together accounting for more than 50% of all FALS cases. Consequently, these genes have formed the basis of most downstream investigations to understand disease pathogenesis (Renton et al., 2014).

Traditionally, gene discovery in FALS relied heavily on family-based linkage studies and candidate gene approaches. However, the advent of next-generation sequencing (NGS) has allowed the collection of genetic data from an unprecedented number of patients, thus facilitating the identification of dozens more genes causing, predisposing to, or associated with ALS (Chia et al., 2018). It is however, incredibly important that new genetic discoveries are given equal scrutiny as those that were initially reported. Alarmingly, some genetic variants in ALS genes are immediately deemed pathogenic upon identification in one patient, without rigorous validation including segregation analysis and/or absence in sufficient numbers of control individuals from relevant populations (Al-Chalabi et al., 2017; Andersen and Al-Chalabi, 2011).

1.4.1.1 *SOD1*

Discovery

In 1991, a collaborative effort used linkage analysis in 18 ALS pedigrees to identify the first ALS-linked locus on the long arm of chromosome 21 (Siddique et al., 1991). This led to the investigation of *SOD1* as a candidate gene, due to its proximity to the microsatellite marker that showed the strongest linkage to disease. A genetic screen of *SOD1* revealed 11 missense mutations segregating with ALS in 13 different ALS families (Rosen, 1993).

ALS mutations

Since the initial report 25 years ago, over 180 different ALS mutations have been reported in *SOD1*, almost all of which are missense mutations with an autosomal dominant pattern of inheritance, though many lack supportive segregation data (Boylan, 2015; Chio et al., 2008; Dion et al., 2009; Renton et al., 2014; Robberecht and Philips, 2013; Sreedharan and Brown, 2013). These mutations are found across most regions of the 153 amino acid SOD1 protein (Taylor et al., 2016). Mutations in *SOD1* account for approximately 12% of familial cases with European ancestry (Paper I; McCann et al., 2017, Boylan, 2015; Renton et al., 2014) and almost 30% of FALS patients with Asian ancestry (Hou et al., 2016; Kwon et al., 2012; Nishiyama et al., 2017). Most *SOD1* mutations are highly penetrant, however reduced penetrance has been observed (Dion et al., 2009). Interestingly, rare recessive *SOD1* mutations have

also been reported, including the p.D90A variant, which was previously considered a benign polymorphism before being identified in a homozygous state in Scandainavian ALS patients (Andersen et al., 1996).

Clinical features

Almost all *SOD1* patients present with classical ALS, generally with limb onset. Frontotemporal impairment is exceptionally rare in these patients (Abel et al., 2012; Andersen and Al-Chalabi, 2011; Boylan, 2015; van Es et al., 2010). However, phenotypic heterogeneity is common both within and between families, including variable age and site of onset and disease duration, (Boylan, 2015). For example, particular *SOD1* mutations have been associated with late age of onset (p.I114T, Paper I; McCann et al., 2017, Al-Chalabi and Hardiman, 2013) and prolonged (p.D90A, Andersen et al., 1996) or rapid (p.A4V, Cudkowicz et al., 1997) disease course.

Function and pathology

The SOD1 protein is ubiquitously expressed in all tissue types and has a highly conserved amino acid sequence across most species (Fridovich, 1995). The main role of SOD1 is to defend against oxygen free radical toxicity (Saccon et al., 2013). SOD1 forms a homodimer upon binding copper and zinc ions, and functions as a dismutase by metabolising superoxide radicals to molecular oxygen and hydrogen peroxide (Saccon et al., 2013).

The pathogenic mechanism by which *SOD1* mutations cause ALS is yet to be established, though it is thought to act through a toxic gain-of-function mechanism (Andersen and Al-Chalabi, 2011). A range of evidence supports this hypothesis. Firstly, overexpression of the mutant SOD1 protein in numerous transgenic animal models results in development of ALS-like phenotypes (Deng et al., 2006; Gurney et al., 1994; Reaume et al., 1996; Wong et al., 1995). Further, there is no correlation between a reduction in SOD1 activity and disease severity, based on analysis in patient derived red blood cell or lymphoblast extracts (Cleveland et al., 1995). Specific mechanisms proposed to mediate the pathogenesis of *SOD1* mutations include excitotoxicity, oxidative stress, ER stress, mitochondrial dysfunction, axonal transport disruption, prion-like propagation, and non-cell autonomous toxicity of neuroglia (Hayashi et al., 2016).

Patients with *SOD1* mutations present with a unique pathology of protein

aggregates in their affected motor neurons that are negative for TDP-43, and instead carry ubiquitinated cytoplasmic SOD1-positive Lewy-body-like hyaline inclusions (Al-Chalabi and Hardiman, 2013; Keller et al., 2012; Mackenzie et al., 2007; Shibata et al., 1996; Tan et al., 2007). This distinct pathology suggests that the mechanism causing *SOD1*-linked ALS is likely different to that underlying the majority of ALS cases that demonstrate TDP-43 pathology (Andersen and Al-Chalabi, 2011).

1.4.1.2 *TARDBP*

Discovery

The *TARDBP* gene encoding TDP-43 was investigated as a candidate ALS gene following the landmark discovery that TDP-43 is the major constituent of the ubiquitinated protein inclusions found in the affected motor neurons of most ALS patients (Neumann et al., 2006). A total of 154 index FALS and 397 SALS patients were screened for genetic variants in *TARDBP*, resulting in the identification of missense variants in two SALS patients, and one FALS patient (Sreedharan et al., 2008). Segregation of the FALS mutation was established when four additional affected family members were found to carry an identical variant (Sreedharan et al., 2008). Subsequently, a genome-wide scan within the extended pedigree confirmed that genetic linkage of disease was restricted to a genomic region encompassing the *TARDBP* gene (Sreedharan et al., 2008). Functional *in vitro* studies supported the role of these mutations in ALS pathogenesis by showing increased fragmentation of mutant TDP-43 compared to wild-type, and further demonstrating that *TARDBP* mutations cause neuronal apoptosis (Sreedharan et al., 2008). Soon after, a second study reported eight additional missense variants in *TARDBP*, three of which were found within ALS families and five in sporadic cases (Kabashi et al., 2008).

ALS mutations

At least 40 mutations have been reported in *TARDBP* among autosomal dominant ALS families and SALS patients, with the vast majority of these found within the C-terminal glycine-rich domain (Chen-Plotkin et al., 2010; Lattante et al., 2013; Robberecht and Philips, 2013; Therrien et al., 2016). Most are missense mutations (Robberecht and Philips, 2013), however some deletions leading to truncated proteins have been reported (Renton et al., 2014; Solski et al., 2012). It is estimated that *TARDBP* mutations account for approximately 3-4% of FALS (Chio et al., 2012; Kabashi et al., 2008; Lattante et al., 2013), and 0.5-2% of SALS (Boylan, 2015; Chio et al., 2012; Kabashi et al., 2008; Lattante et al., 2013), although

geographical variation is apparent, such as the high frequency of the p.A382T mutation in Sardinia (Boylan, 2015; Chio et al., 2011a; Lattante et al., 2013; Renton et al., 2014). *TARDBP* mutations have been identified in patients of various ancestries.

Clinical features

Classical ALS with limb onset is generally observed among *TARDBP* mutation carriers, with some reports of extended survival compared to SALS patients who are negative for known ALS mutations. Some rare *TARDBP* patients have reportedly experienced symptoms of FTD or Parkinson's disease (Boylan, 2015; Corcia et al., 2012; Lattante et al., 2013).

Function and pathology

TDP-43 is an ubiquitously expressed RNA-binding protein, closely resembling the family of heterogeneous nuclear ribonucleoproteins (hnRNPs). It plays roles in transcriptional repression and activation, mRNA splicing and nucleo-cytoplasmic RNA transport (Chen-Plotkin et al., 2010; Wang et al., 2008; Warraich et al., 2010).

The ALS pathogenic mechanism induced by *TARDBP* mutations is still heavily debated (Feneberg et al., 2018). A loss of function mechanism is supported by the development of ALS relevant pathology in full and partial TDP-43 knockdown *in vitro* (Schwenk et al., 2016) and *in vivo* (Schmid et al., 2013; Wu et al., 2012b) models. For example, TDP-43 knockdown in cell models caused impairments in RNA-binding capacity and splicing activities (Schwenk et al., 2016), while progressive motor dysfunction has been induced by selective and ubiquitous silencing of *TARDBP* in mice (Wu et al., 2012b) and zebrafish (Schmid et al., 2013), respectively. On the other hand, mutant TDP-43 overexpression in animals and cell lines has led to a range of ALS specific changes such as mislocalisation and nuclear clearance of endogenous TDP-43, phosphorylation of TDP-43 and formation of ubiquitinated TDP-43 positive aggregates (Igaz et al., 2009; Kabashi et al., 2010; Nonaka et al., 2009).

Though it remains to be seen whether TDP-43 aggregates are a cause or consequence of disease, the underlying mechanism of aggregate formation is certainly significant for a better understanding of ALS pathogenesis. As discussed in Section 1.3.5.1, the glycine-rich C-terminal domain of TDP-43 is suspected to play an integral role in the contribution of this protein to ALS pathogenesis. This domain has the ability to act as a prion-like domain, which may act as a template to induce conversion of natively folded proteins and cause entrapment and aggregation (Kim et al., 2013;

Robberecht and Philips, 2013). Induction of aggregate formation by prion-like domains is further supported by the presence of prion-like domains in other ALS proteins, including FUS, TAF15, hnRNPA1, hnRNPA2/B1 and EWSR1.

The ALS hallmark TDP-43 positive protein aggregates observed in the cytoplasm and glia of affected motor neurons (described in Section 1.3.4) are observed in 98% of all ALS patients, including those with causal mutations in *TARDBP* (Chen-Plotkin et al., 2010; Feneberg et al., 2018; Lomen-Hoerth et al., 2002; Van Deerlin et al., 2008). Further, it has been suggested the TDP-43 positive inclusions may actually be more abundant in patients with *TARDBP* mutations compared to other ALS patients (Van Deerlin et al., 2008).

1.4.1.3 *FUS*

Discovery

Soon after ALS mutations were identified in *TARDBP*, two groups independently investigated *FUS* as a candidate ALS gene. An ALS-linked locus on the long arm of chromosome 16 was initially reported by Ruddy et al. (2003), and was further refined to a genomic region containing 400 genes by Vance et al. (2009). Following investigation of six candidate genes, a single point mutation was identified in *FUS* (Vance et al., 2009). This mutation segregated with disease in all six affected family members. A further four families also carried this mutation, and two additional *FUS* mutations were found within unrelated families and probands (Vance et al., 2009). Kwiatkowski et al. (2009), used loss of heterozygosity (LOH) mapping in a consanguineous family, to also link disease to chromosome 16. Subsequent screening of *FUS* identified a novel homozygous mutation, while another two *FUS* mutations were found in two more families whom also showed genetic linkage to chromosome 16 (Kwiatkowski et al., 2009). Additional FALS patient screening revealed more novel variation in *FUS* and a total of 13 distinct *FUS* mutations in 17 ALS kindreds (Kwiatkowski et al., 2009).

ALS mutations

To-date, at least 58 mutations in *FUS* have been identified (Deng et al., 2014; Lattante et al., 2013). Interestingly, residue 521 of the FUS protein is the most frequently mutated FUS residue among ALS patients, with five different amino acid substitutions reported at this position. Similar to *TARDBP*, the majority of *FUS* validated mutations occur in the glycine-rich RNA-binding C-terminal domain of the protein (Blair et al., 2010; Deng et al., 2014; Renton et al., 2014). While most reported

FUS mutations are autosomal dominant point mutations (including missense and splicing), some structural variations have also been reported (Boylan, 2015; Chio et al., 2009b; Conte et al., 2012; Deng et al., 2014; Lattante et al., 2013; Zou et al., 2013). Recessive inheritance has also been observed for *FUS* (Kwiatkowski et al., 2009). Approximately 4% and 1% of familial and sporadic patients carry a *FUS* mutation respectively (Boylan, 2015; Deng et al., 2014; Renton et al., 2014). Interestingly, a number of *FUS* mutations have been reported to occur *de novo* (Chio et al., 2011b; DeJesus-Hernandez et al., 2010; Zou et al., 2013).

Clinical features

FUS mutations are commonly associated with more aggressive forms of ALS, including juvenile ALS (Andersen and Al-Chalabi, 2011; Conte et al., 2012; Zou et al., 2013). *FUS* mutations are also associated with bulbar onset, early onset and a rapid disease course (Paper I; McCann et al., 2017, Lattante et al., 2013; Millecamps et al., 2010). Cognitive impairment is also observed in rare *FUS* cases (Blair et al., 2010; Deng et al., 2014; Lattante et al., 2013).

Function and pathology

FUS belongs to the FET protein family of highly conserved RNA-binding proteins that also includes EWS and TAF15 (Tan and Manley, 2009). FUS is a predominately nuclear protein, but shuttles between the nucleus and cytoplasm (Zinszner et al., 1997). It contains multiple protein domains, including an N-terminal transcriptional activation domain rich in serine, tyrosine, glycine and glutamine (SYGQ); a RNA-recognition motif (RRM); three arginine glycine glycine repeat regions (RGG1-3); a zinc-finger motif; and a highly conserved non-classical nuclear localisation signal domain located in the C-terminus (Deng et al., 2014). FUS targets thousands of RNA molecules by binding through its RRM domain (Daigle et al., 2013). It plays a role in RNA transcription, splicing, transport and processing (Boylan, 2015; Deng et al., 2014; Yang et al., 2010).

FUS mediated toxicity is thought to be related to factors including its propensity for stress granule formation, its prion-like domain and arginine methylation (Deng et al., 2014). *FUS* knockdown mouse models do not show any ALS related abnormalities (Kino et al., 2015; Sharma et al., 2016), suggesting that FUS is not acting through a loss-of-function pathogenic mechanism. However, transgenic mouse models carrying ALS-linked *FUS* mutations have been shown to develop progressive motor neuron degeneration, implicating a toxic gain-of-function (Sharma et al., 2016). The

toxic activity of mutant FUS is at least in part mediated by its ability to bind RNA, as deletion of the RRM domain renders the protein incapable of causing neurological defects (Daigle et al., 2013). Further, mutant FUS is prone to mislocalisation to the cytoplasm (Dormann et al., 2010), where it is exposed to a unique set of RNA substrates, potentially leading to toxic interactions. It has also been shown that mutant FUS has stronger affinity for the survival of motor neuron (SMN) protein (implicated in spinal muscular atrophy) than wild-type FUS (Chari et al., 2009) and affects SMN related spliceosome activity (Sun et al., 2015) and transport to axons (Groen et al., 2013). As described for *TARDBP* (Section 1.4.1.2), *FUS* mutants are also thought to promote aggregation through the prion-like domain of their aberrant protein products (Kim et al., 2013; Robberecht and Philips, 2013).

Post-mortem studies suggest that motor neuron loss in FUS patients is most extensive in the spinal cord and brain stem, and less pronounced in the motor cortex (Deng et al., 2014). Unlike most ALS patients, *FUS* patients have ubiquitinated protein aggregates that are negative for TDP-43, though positive for FUS (Kwiatkowski et al., 2009; Vance et al., 2009). Another interesting observation is the presence of basophilic inclusions in *FUS* p.R525L patients, that appear to be absent from other *FUS* mutation carriers, though variably observed among other ALS patient subsets (reviewed by Deng et al., 2014). These observations, together with links between FUS and stress granules (Vance et al., 2013), suggest that a gain-of-function pathogenic mechanism is most likely underlying *FUS* mediated toxicity.

1.4.1.4 *C9orf72*

Discovery

Between 2007 and 2011, multiple genetic linkage studies of kindreds with inheritance of ALS, ALS/FTD and FTD identified a locus on the short arm of chromosome 9 (Boxer et al., 2011; Gijselinck et al., 2010; Le Ber et al., 2009; Luty et al., 2008; Morita et al., 2006; Pearson et al., 2011; Valdmanis et al., 2007; Vance et al., 2006). Genome-wide association studies (GWAS) provided further support for this locus (Laaksovirta et al., 2010; Shatunov et al., 2010; Van Deerlin et al., 2010; van Es et al., 2009). Analysis in the Finnish population (Laaksovirta et al., 2010) refined the locus to a 232kb linkage disequilibrium block comprising a 42 SNP risk-haplotype that was later found to be shared by ALS patients linked to this locus, in populations with European ancestry (Mok et al., 2012). Two independent groups concurrently reported the identification of a polymorphic GGGGCC hexanucleotide repeat located between exons 1a and 1b

of *C9orf72* (DeJesus-Hernandez et al., 2011; Renton et al., 2011). The first did so using deep sequencing of the disease linked region within affected families (Renton et al., 2011), while the other group conducted haplotype analysis of the intronic region of *C9orf72* in which the expansion was found to lie (DeJesus-Hernandez et al., 2011).

ALS mutations

Expansion of the hexanucleotide repeat in *C9orf72* is the most common known cause of ALS, accounting for up to 40% of FALS cases and 5-10% of SALS patients with European ancestry (Boylan, 2015; Majounie et al., 2012; Renton et al., 2014), though it is rare among other patient populations (Majounie et al., 2012; Nishiyama et al., 2017). Interestingly, SALS patients who carry the expansion also have the aforementioned 42 SNP risk founder haplotype, suggesting that either some cases are misclassified FALS patients and/or the mutation is not always completely penetrant (discussed further in Section 1.6.1). The *C9orf72* hexanucleotide repeat is highly polymorphic, with neurologically normal individuals typically carrying anywhere up to 20 repeat units (Ng and Tan, 2017), though larger repeats of up to 32 repeat units have been observed in rare control individuals (Theuns et al., 2014; van der Zee et al., 2013). Pathogenic repeats cause autosomal dominant inheritance of disease, and pathogenic alleles are thought to contain more than 30 repeat units (Renton et al., 2011). However, this number is largely debated, and the number of repeats required to initiate disease onset remains to be determined (Ng and Tan, 2017). Importantly, patients with thousands of repeat units have been identified using Southern blotting (Dols-Icardo et al., 2014). The threshold of 30 repeats is largely a result of the inability of the repeat primed PCR method (the main technique routinely used to analyse the expansion) to accurately size expansions larger than this (Dols-Icardo et al., 2014; Ng and Tan, 2017). Southern blotting techniques have however detected up to 4,500 repeat units in ALS patients, and data suggest that ALS patients harbour larger repeat expansions than FTD patients (Dols-Icardo et al., 2014).

Clinical features

Patients carrying the expansion can present with the pure form of either ALS or FTD, or with co-morbid ALS/FTD (Byrne et al., 2012b; DeJesus-Hernandez et al., 2011; Majounie et al., 2012; Renton et al., 2011; van Rheenen et al., 2012). Interestingly, many expansion positive ALS patients also exhibit cognitive deficits, without meeting the criteria for an FTD diagnosis (Byrne et al., 2012b; van Rheenen et al., 2012). In rare cases, Parkinson's and Huntington's phenotypes are also seen (O'Dowd et al., 2012; van Rheenen et al., 2012). The disease penetrance of the expansion seems to be

zero in persons under 35 years of age, reaching 50% penetrance at approximately 60 years, and full penetrance at approximately 85 years (Paper I; McCann et al., 2017 Majounie et al., 2012; Williams et al., 2013). A popular hypothesis is that repeat expansion size correlates with the phenotypic features of disease including age of onset, disease progression and presence of cognitive deficits, however as yet there is no consensus as to whether this is true, as no such correlations have yet been identified (Dols-Icardo et al., 2014; Gijselinck et al., 2016; Ng and Tan, 2017).

Function and pathology

The C9orf72 protein shows structural similarity to DENN (differentially expressed in normal and neoplasia) proteins (Burrell et al., 2016). C9orf72 plays roles in nuclear and endosomal membrane trafficking, actin dynamics and autophagy (Brown and Al-Chalabi, 2017; Farg et al., 2014; Sivadasan et al., 2016). *C9orf72* is transcribed into three major transcripts, which encode the two protein isoforms, C9orf72 a and b (Farg et al., 2014). Falling between exons 1a and 1b, the hexanucleotide repeat region forms part of the functional core promoter, driving expression of all three transcripts (Gijselinck et al., 2012).

Three major mechanisms have been proposed to underlie the pathogenicity of hexanucleotide repeat expansions in *C9orf72*. A gain-of-function toxicity is the favoured hypothesis, owing to the genetic dominance of the expansions and the absence of disease in individuals carrying null alleles or missense variants (Taylor et al., 2016).

Haploinsufficiency of the C9orf72 protein has also been postulated as a mechanism of action based on observations of reduced levels of C9orf72 mRNA in patient tissues (Belzil et al., 2013; DeJesus-Hernandez et al., 2011; Gijselinck et al., 2012) as well as induced pluripotent stem cell derived human motor neurons (Almeida et al., 2013) and zebrafish (Ciura et al., 2013) models. Reduced expression is likely mediated through epigenetic mechanisms (Belzil et al., 2013; Gendron et al., 2014). However, the potential contribution of *C9orf72* haploinsufficiency to disease pathogenesis is unclear, as how the observed reduction in C9orf72 mRNA levels correlate with protein levels in patients remains unknown (Gendron et al., 2014), while animal models with reduced *C9orf72* expression show variable phenotypes (Burrell et al., 2016; Gendron et al., 2014; Mizielinska et al., 2014; Taylor et al., 2016).

RNA foci containing both sense (GGGGCC) and antisense (CCCCGG) repeat

RNA transcripts have been shown to accumulate in neuronal tissue from expansion carriers (DeJesus-Hernandez et al., 2011). In other neurodegenerative diseases, similar RNA foci have been shown to cause defects in RNA splicing by sequestration of RNA-binding proteins to elicit pathogenic effects (La Spada and Taylor, 2010). Further, RNA-binding proteins with affinity for these repeat sequences have been observed to co-localise with RNA foci in affected patient tissues (Lee et al., 2013; Xu et al., 2013).

DPRs produced by unconventional translation have also been proposed to mediate *C9orf72* pathogenicity. RAN translation was first observed in spinocerebellar ataxia type 8, caused by a repeat expansion in *ATXN8OS* (Zu et al., 2011). This led to the investigation of RAN translation of expanded GGGGCC repeats in *C9orf72*, which revealed RAN translation of both the sense and antisense strand of the expansion in all six reading frames, facilitating the generation of five distinct DPRs (Mori et al., 2013b; Zu et al., 2013). *In vitro* and *in vivo* evidence indicates that these DPRs are toxic, forming neuronal cytoplasmic and intranuclear inclusions in affected motor neurons of the cerebellum, and frontal and temporal lobes (Ash et al., 2013; Gendron et al., 2013; Mori et al., 2013a,b; Zu et al., 2013). Animal models have demonstrated that DPR toxicity does appear to induce motor defects (Mizielinska et al., 2014; Ohki et al., 2017; Swaminathan et al., 2018). It has been demonstrated that the toxicity of these DPRs is largely attributable to those which contain arginine (Freibaum et al., 2015; Mizielinska and Isaacs, 2014).

The pathology observed in *C9orf72* cases is typical of most ALS patients in that they carry ubiquitin- and TDP-43-positive protein inclusions (Boylan, 2015; DeJesus-Hernandez et al., 2011; Mackenzie et al., 2013; Renton et al., 2011). However, many *C9orf72* patients also carry additional star shaped DPR containing protein inclusions in the cerebellum and frontal and temporal lobes, though these are noticeably absent from the spinal cord (Boylan, 2015; DeJesus-Hernandez et al., 2011; Mackenzie et al., 2013; Renton et al., 2011).

1.4.1.5 *UBQLN2*

Discovery

Deng et al. (2011) identified a five-generation pedigree with 19 ALS patients who exhibited dominant inheritance of disease, with reduced penetrance in females. After eliminating known ALS genes, a genome-wide linkage analysis was performed using autosomal markers but failed to identify a disease-linked locus. Due to the

observed lack of male-to-male transmission, linkage analysis of the X chromosome was subsequently conducted, revealing a distinct linkage peak. Detailed mapping refined the disease-linked locus to a region encompassing 191 protein-encoding genes, 41 of which were sequenced as candidates. This revealed a unique missense mutation in *UBQLN2*. Four additional *UBQLN2* mutations were subsequently identified from 188 ALS families negative for known ALS genes and lacking male-to-male transmission, two of which were supported by segregation, while the other two were found in probands. Soon after, *UBQLN2* mutations were confirmed as a cause of ALS, when our laboratory used whole-exome sequencing (WES) to identify another novel missense mutation, present in two multi-generational, apparently unrelated ALS families (Williams et al., 2012b).

ALS mutations

X-linked dominant mutations in *UBQLN2* account for approximately 1% of FALS patients (Boylan, 2015). At least 16 missense *UBQLN2* mutations have been reported in FALS, SALS and ALS/FTD patients with varied ancestries including Australian, German, Turkish, Italian, American and French-Canadian (Daoud et al., 2012; Deng et al., 2011; Fahed et al., 2014; Gellera et al., 2013; Ozoguz et al., 2015; Synofzik et al., 2012; Williams et al., 2012b). Studies have found *UBQLN2* mutations to be absent from ALS/FTD patient populations from Korea, the Netherlands, France and Ireland (Kim et al., 2014; McLaughlin et al., 2014; Millecamps et al., 2012; van Doormaal et al., 2012). Most *UBQLN2* mutations identified to-date have been reported within the proline-rich repeat region of the protein (Deng et al., 2011). Interestingly, a p.P506S mutation was identified in a large kindred where both males and females were affected. This family displayed multiple phenotypes including ALS/FTD, spastic paraplegia, bulbar palsy and multiple sclerosis (Vengoechea et al., 2013).

Clinical features

Most patients with a *UBQLN2* mutation have an ALS phenotype, though some do go on to develop co-morbid FTD or more mild cognitive deficits (Deng et al., 2011; Renton et al., 2014; Synofzik et al., 2012; Vengoechea et al., 2013; Williams et al., 2012b). Varied clinical presentation has been observed among *UBQLN2* patients, including earlier onset in male compared to female mutation carriers (Deng et al., 2011; Williams et al., 2012b). Mutations in this gene generally show quite high disease penetrance (Williams et al., 2012b).

Function and pathology

The UBQLN2 protein belongs to the ubiquilin protein family, which is involved in proteasome-mediated protein degradation. These proteins are characterised by their ubiquitin-associated (UBA) and ubiquitin-like (UBL) domains that mediate their degradation activity (Rothenberg et al., 2010). A key function of the UBQLN2 protein is to recruit autophagosomes to polyubiquitinated aggregates through interactions involving its UBA domain (Nguyen et al., 2018a). Importantly, within the UBQLN2 protein, this UBA domain sits next to a PXX (proline-X-X aminio acid sequence) domain, thought to be important for protein-protein interactions (Aitio et al., 2010; Kleijnen et al., 2000). As previously mentioned, this domain harbours many disease-causing mutations (Daoud et al., 2012; Gellera et al., 2013; Williams et al., 2012b).

UBQLN2 mutations are thought to take pathogenic effect through impairment of the UPS and/or autophagic dysfunction. A study of neuronal cells overexpressing mutant UBQLN2 showed accumulation of poly-ubiquitinated proteins leading to inclusion-body formation, and also reduced co-localisation between the UBQLN2 protein and an essential autophagosome-lysosome fusion factor, ATG9/ATG16L1 (Osaka et al., 2016). Mutant forms of UBQLN2 have also been shown to impair endosomal pathways. A study of a cellular model expressing the ALS mutation p.E478G, showed inhibition of endosomal vesicle formation and trafficking, and increased formation of inclusion bodies (Osaka et al., 2015). Further, a rat model with a p.P497H ALS mutation showed loss of motor neurons and reduced levels of the early endosome antigen 1, indicating impaired endosomal function, which may underlie motor neuron loss (Wu et al., 2015).

The TDP-43 positive protein aggregates that are found in the motor neurons of the majority of ALS patients, including those carrying *UBQLN2* mutations, are also immunoreactive for the UBQLN2 protein (Boylan, 2015; Renton et al., 2014; Williams et al., 2012b). Williams et al. (2012b) observed compact and skein-like inclusions in spinal cord tissue from a *UBQLN2* mutation carrier, and showed these inclusions also contained ubiquitin, TDP-43 and FUS. Axonal loss in the corticospinal tract, loss of anterior horn cells and astrocytosis has also been reported in post-mortem spinal cord tissue from *UBQLN2* mutation carriers (Deng et al., 2011).

1.4.1.6 Other ALS genes

A number of less common ALS genes have been identified through linkage analysis of large families, and subsequent candidate gene screening. Such genes include *SETX* (Chen et al., 2004), *ALS2* (Hadano et al., 2001), *DCTN1* (Puls et al., 2003), *VAPB* (Nishimura et al., 2004) and *ANG* (Greenway et al., 2004, 2006). Candidate gene approaches without linkage analysis in cohorts of smaller ALS kindreds revealed mutations that cause ALS in *FIG4* (Chow et al., 2009), *SQSTM1* (Fecto et al., 2011) and *GLE1* (Kaneb et al., 2015).

A rapid rise in the number of genes implicated in the aetiology of ALS was seen following the widespread adoption of NGS technologies. In fact, this rate is now so rapid that there appears to be a doubling of the number of reported ALS genes every four years (Al-Chalabi et al., 2017). Using SNP chip genotyping technology and homozygosity mapping, the *OPTN* gene encoding optineurin was found to harbour autosomal recessive mutations causing ALS in Japanese kindreds (Maruyama et al., 2010). Family-based studies utilising WES have facilitated the discovery of multiple ALS genes including *VCP* (Johnson et al., 2010), *PFN1* (Wu et al., 2012a), *HNRNP* genes (Kim et al., 2013), *MATR3* (Johnson et al., 2014b), *CHCHD10* (Bannwarth et al., 2014), *CCNF* (Williams et al., 2016b) and *TIA1* (Mackenzie et al., 2017). The family based approach was also successfully employed using whole-genome sequencing (WGS) to implicate *ERBB4* as an ALS gene (Takahashi et al., 2013). More recently, NGS has been used for gene burden analysis in larger cohorts of familial probands. Using this strategy, *TUBA4A* (Smith et al., 2014), *TBK1* (Cirulli et al., 2015; Freischmidt et al., 2015) and *NEK1* (Kenna et al., 2016) have been implicated in the aetiology of ALS.

While this expansion of the ALS gene spectrum is exciting, it is important to note that mutations in each of these less common ALS genes have been reported in 1% or less of familial cases, and only rarely in sporadic cases. Nevertheless, many of these genes cluster together in biological pathways and processes, implicating mechanisms of disease that may be common to all ALS patients. These pathways include RNA processing, protein homeostasis and degradation as well as vesicular trafficking (discussed in Section 1.3.5). As such, these rare mutations offer the opportunity for the generation of various disease models to investigate the mechanisms that are widespread in ALS pathogenesis. Gene discovery has provided targets for downstream research into the cellular and functional defects that contribute to the onset and progression of ALS. While potential gene therapy or screening may only be of use

to the patients and families directly affected by these mutations, every ALS patient that can be treated and every family in which disease can be prevented in the next generation is of paramount importance to fighting ALS.

1.4.2 Sporadic ALS

Sporadic ALS patients have no known relatives whom have been diagnosed with ALS, and this often causes the misconception that there is no genetic component underlying disease in these patients (Taylor et al., 2016). However, there may be some degree of genetic predisposition underlying SALS. Indeed, many gene mutations identified in FALS have subsequently been found in SALS, suggesting that some SALS patients may be familial cases with incomplete penetrance, or may simply have insufficient family history available for an accurate classification to be determined. Between 1-3% of sporadic patients carry a mutation in the *SOD1* gene (Gamez et al., 2006), while approximately 5% have an expansion in *C9orf72* (Renton et al., 2014). Rare mutations in other ALS genes including *TARDBP*, *FUS*, *HNRNPA1*, *SQSTM1*, *VCP*, *OPTN* and *PFN1* have also been reported in SALS patients (Taylor et al., 2016).

Heritability studies suggest a genetic component underlies approximately 60% of sporadic cases (Al-Chalabi et al., 2010; McLaughlin et al., 2015). However, the genetic architecture potentially contributing to SALS is more complex than the simple autosomal dominant inheritance observed in most FALS patients. There may be a small number of genetic variants, each conferring a moderate disease-risk, with the sum of risk equating to disease onset (Al-Chalabi et al., 2017). Alternatively, a large number of common variants may each marginally confer a small disease-risk (Al-Chalabi et al., 2017). Indeed, a complex combination of rare and common variants may underlie SALS and still others may also act as modifiers of clinical phenotypes (Al-Chalabi et al., 2017; Taylor et al., 2016). It is also likely that various environmental factors interact with genetic variation in order to cause ALS, or influence its progression.

1.4.2.1 Genetic association with disease

GWAS have been a widely adopted strategy for identifying risk loci for a range of diseases. GWAS is based on the principle of "common disease - common variant", whereby the accumulation of small contributions of many common variants results in disease (Al-Chalabi et al., 2017). Although ALS is not a common disease, the high

percentage of sporadic cases suggests that isolated mutations of large effect are not the only cause of ALS, therefore it is likely that common genetic variation may contribute to disease-risk (Al-Chalabi et al., 2017).

GWAS have provided considerable insight into ALS, most notably by its use in the identification of the association between ALS and the short arm of chromosome 9, where the pathogenic expansion in *C9orf72* was later found (Laaksovirta et al., 2010; Shatunov et al., 2010; van Es et al., 2009). Known SNPs have also been reported as risk loci for SALS, with the genes *UNC13A* (Shatunov et al., 2010; van Es et al., 2009), *C21orf2* (van Rheenen et al., 2016) and *GPX3-TNIP1* (Benyamin et al., 2017) being those supported by the most robust evidence. However, many other SALS-based GWAS have identified potential risk loci that have failed to be independently replicated (Renton et al., 2014). It is possible, and even likely, that many risk variants are population specific, deeming it difficult to obtain sufficient sample numbers to first identify association and then to replicate the findings (Al-Chalabi et al., 2017). This highlights a caveat of ALS GWAS studies, which have largely focused on European-based populations (Al-Chalabi et al., 2017; Renton et al., 2014). The contribution of any GWAS risk-loci to disease in more ancestrally diverse cohorts is questionable, and requires further validation (Renton et al., 2014).

A number of GWAS have also been performed to identify phenotype-modifying variants. Survival has been linked to variation in the genes *KIFAP3* (Landers et al., 2009), *EPHA4* (Van Hoecke et al., 2012) and *UNC13A* (Gaastra et al., 2016), though these results remain to be replicated (Renton et al., 2014). Looking beyond GWAS, candidate genes such as textitATXN2 have also been associated with SALS Elden et al. (2010).

1.4.2.2 Genetic burden

Gene-based burden testing is an increasingly useful approach for identifying genes involved in disease. In such an analysis, the cumulative frequency of "qualifying variants" meeting a given criteria is compared between cases and controls, to establish the burden of variants in a gene (Guo et al., 2016). Qualifying variants (eg. rare non-synonymous variants) are those that are more likely pathogenic, and can be defined by various filters, such as minor allele frequency (MAF), functional consequences and *in silico* protein predictions (Guo et al., 2016).

A landmark genetic burden analysis in ALS was reported by Cirulli et al. (2015), which assessed genetic burden in each of the known ALS genes. No single gene was found to contribute more than 1% to SALS patients and many genes known to segregate with disease did not reach significance. Two novel genes, *TBK1* and *NEK1*, were implicated in SALS, contributing to 0.9% and 0.7% respectively. *NEK1* was also independently identified in another gene based burden analysis, which specifically implicated the p.R261H variant (Kenna et al., 2016). Additionally, Smith et al. (2014) found that *TUBA4A* carried a genetic burden of rare and predicted damaging variants among SALS patients, second only to *SOD1*. There have also been reports suggesting that patients with co-morbid ALS/FTD may also carry a high genetic burden in known ALS genes (Dols-Icardo et al., 2018).

1.5 Approaches for gene discovery

The widespread adoption of NGS has lead to an explosion of available sequencing data. As a result, genetic analysis techniques have had to rapidly evolve to effectively utilise this volume of data. By combining NGS data with the following analysis techniques, our power for identifying novel disease genes has increased. In the later half of the twentieth century at the time of the Human Genome Project, linkage analysis and positional cloning techniques dominated genetic research. Many disease genes were identified within large families with Mendelian inheritance patterns, including *SOD1* in ALS. Today, NGS data is widely used for genetic analysis. WGS, WES and targeted sequencing are powerful tools for familial disease gene discovery (Ott et al., 2015), and also represent opportunities to better understand and appreciate common genetic variation.

1.5.1 Linkage analysis

Genetic recombination occurs when homologous chromosomes participate in random crossover events to facilitate the exchange of DNA segments. These crossover events are less likely to separate DNA sequences that lie close together on a chromosome. The principle of genetic linkage is that DNA sequences in close proximity to each other are more likely to remain together after these recombination events, and therefore are more likely to be inherited together (Pulst, 1999; Williams, 2018). Thus, linkage analysis identifies chromosomal regions that are co-inherited with a phenotype, usually

disease affection status.

To perform linkage analysis, highly polymorphic genetic markers are required to trace segregation. Genetic markers used in linkage analysis include microsatellites and single nucleotide polymorphisms (SNPs). Microsatellites are short tandem repeats, consisting of a variable number of di-, tri- or tetra-nucleotide repeat units that are multi-allelic and highly variable between individuals. These properties increase the likelihood of a heterozygous genotype in any given individual, which means maternal and paternal alleles can usually be distinguished (Pulst, 1999). Being multi-allelic, microsatellites are highly informative genetic markers, providing insights into parental origins (Dewoody and Dewoody, 2005; Pulst, 1999).

SNPs are single nucleotide polymorphisms in the genome that are usually bi-allelic and common throughout the population. They represent the most common form of genetic variation and are easily detectable using high-throughput and automated genotyping (Dewoody and Dewoody, 2005; Ott et al., 2015; Pulst, 1999). While not as informative as microsatellites, SNPs are useful markers for linkage analysis owing to the ability to genotype hundreds of thousands of SNPs in many people with high efficiency.

Genome-wide linkage analysis uses microsatellite or SNP markers scattered throughout the genome to identify those that co-segregate with disease within a family, and therefore define a chromosomal region linked to disease. When using microsatellite markers, between 300 and 400 sites, spaced out by an average of 5-15 centimorgans, are typically assessed (Borecki and Province, 2008). When using a generic SNP micoarray, such as the Infinium CoreExome-24 BeadChip (Illumina), over 500,000 markers from different genomic regions (missense, nonsense and synonymous exonic variants, as well as intronic splicing or promoter variants), can be genotyped. As SNP arrays interrogate hundreds of thousands of genomic sites, and are amenable for use in large cohorts in a high-throughput fashion, they have become a widely adopted technology (Ott et al., 2015). Genome-wide linkage analysis is an immensely powerful approach and importantly, is unbiased, in that it does not rely on any prior hypothesis of the locus responsible for the genetic basis for disease.

Using statistical models, family-based linkage analysis uses a family pedigree and marker genotypes from informative family members to calculate the likelihood of each marker co-segregating with disease due to linkage or purely by chance (Altshuler et al., 2008; Pulst, 1999). The resulting numeric measure for each genetic marker is referred

to as the logarithm of odds (LOD) score, a concept developed by Morton (1955). For a marker to reach significance and be considered disease-linked, a LOD score of at least 3.3 is required (in a genome-wide scan), indicating an odds ratio of greater than 1000:1 that the marker is linked to disease. A LOD score of less than -2 is conversely evidence against linkage. Markers with intermediate LOD scores remain ambiguous (Lander and Kruglyak, 1995; Pulst, 1999). In cases where statistically significant linkage is not met anywhere in the genome for a single pedigree, multiple families with the same disease can be summed together to strengthen the linkage signal. This approach is based on the assumption that the same disease locus is common to the summed families. The genomic region over which the LOD peak lies represents the disease-linked locus, and sequencing of candidate genes contained within this region has frequently revealed disease causal mutations. Family-based linkage analysis has been tremendously successful for diseases showing complete penetrance and autosomal dominant or recessive Mendelian inheritance, increasing the number of known disease genes from just 100 in the mid 1980s to over 2,000 by the late 2000s (Altshuler et al., 2008).

However, there are a number of factors that may confound linkage analysis, and accounting for these factors is vital to obtain accurate genetic linkage results (Kruglyak et al., 1996). Incomplete disease penetrance hinders the identification of genetic linkage, and as previously discussed is a prominent feature of many ALS pedigrees. To overcome this, parametric linkage analysis can incorporate liability classes to inform the statistical model of the likelihood that an individual carrying the causal mutation will be affected by disease at a given age. Liability classes can also be used to specify the effect of sex on disease state. It is also necessary to provide expected disease allele frequency to the model, to describe how frequently the disease allele is likely to be seen in a population (Kruglyak et al., 1996).

1.5.2 Next-generation sequencing (NGS)

Sanger sequencing has dominated genetic research for the last 30 years and remains the gold standard for validating genetic variants. However, Sanger sequencing is not amenable to high-throughput applications, such as sequencing many target genes through large cohorts in parallel (Moller et al., 2015). NGS is an umbrella term for a number of sequencing approaches, with their underlying commonality being the sequencing of multiplex libraries containing millions of DNA fragments in a massively

parallel way, leading to megabases of DNA sequence output (Ng et al., 2009). The use of these technologies has led to an explosion of available DNA sequence data, and this has consequently demanded the development of sophisticated bioinformatics strategies to gain meaning from this data. Standard bioinformatics processing includes quality control checks, alignment to the reference genome and variant calling. Many different software tools now exist for each of these processing steps, each with innate advantages and disadvantages.

NGS technologies and the huge amounts of available genomic data they produce have changed the scope of genetic research. This is evidenced by a drastic increase in the volume of genetic discoveries since the widespread adoption of these technologies. The number of identified disease genes has jumped from approximately 2,000 in 2007 to almost 5,000 in 2017, while numerous disease associated genes, *de novo* mutations and oligogenic disease factors have also been uncovered (Fernandez-Marmiesse et al., 2018). These technologies have also hugely expanded the catalogue of common genetic variation across the globe, most notably seen in publicly available control databases such as dbSNP (`https://www.ncbi.nlm.nih.gov/SNP/`), ExAC (Exome aggregation consortium, Lek et al., 2016) and gnomAD (Genome aggregation database, Lek et al., 2016), which are now essential resources for a plethora of medical research applications.

1.5.2.1 Whole-exome sequencing (WES)

The exome refers to the protein-coding regions (exons) of all the genes in the genome, which equates to approximately 180,000 exons from 20,000 genes, and 35 megabases of DNA sequence, representing just 1-2% of the human genome (Liu et al., 2015; Moller et al., 2015). Though this may seem like a small proportion, an estimated 85% of all identified disease causal mutations lie in protein coding exons (Liu et al., 2015). Further, the exome represents the best understood genomic region, and therefore the influence of variation in the exome is most easily interpreted. Compared with WGS, the amount of data produced by WES provides a more time and cost effective pipeline (Fernandez-Marmiesse et al., 2018; Lelieveld et al., 2016). Further, following the widespread adoption of WES in genomic research, established workflows can facilitate effective and accurate data analysis and management.

Since becoming widely adopted in 2009, WES has significantly contributed to novel disease gene discoveries, particularly in neurodegenerative disease research (Liu et al., 2015). In 2010, the first successful applications of WES were reported.

Four unrelated patients suffering from Miller syndrome underwent WES and control filtering, resulting in the identification of a single candidate gene, *DHODH*, which encodes a pivotal enzyme in the pyrimidine *de novo* biosynthesis pathway (Ng et al., 2010b). The mutation was confirmed by Sanger sequencing, and further identified in three additional patients. In the same year, WES was also employed to discover disease genes for Kabuki (*MLL2*; Ng et al., 2010a), Schinzel-Giedion (*SETBP1*; Hoischen et al., 2010), and Sensenbrenner (*WDR35*; Gilissen et al., 2010) syndromes.

WES has proven especially powerful when combined with traditional gene mapping approaches such as genetic linkage analysis (Liu et al., 2015). In cases where linkage analysis was able to identify the disease-linked loci, but candidate gene screening yielded inconclusive results, WES has often been able to reveal the causal mutation. For instance, the disease locus for spinocerebellar ataxia-22 was mapped to chromosome 1p21-q23 in the early 2000s (Chung et al., 2003; Verbeek et al., 2002), however the causal mutation was not identified until 2012, after WES was used to screen the disease-linked region (Lee et al., 2012).

In addition to its use in accelerating gene discovery, WES also provides a valuable diagnostic tool. Recent studies have consistently shown that diagnosis rates for patients with previously undiagnosed, but suspected genetic conditions, sits at approximately 25% (Farwell et al., 2015; Gahl et al., 2012; Lee et al., 2014; Sawyer et al., 2016; Yang et al., 2013, 2014). This figure is estimated to be a 50% improvement on that previously achieved using traditional Sanger sequencing methods for diagnosis (Neveling et al., 2013a).

1.5.2.2 Whole-genome sequencing (WGS)

WGS provides an unbiased NGS approach by sequencing the entire human genome including coding, untranslated, miRNA, promoter, repressor/enhancer, intronic and intergenic regions (Stranneheim and Wedell, 2016). By covering the entire genome, WGS offers the opportunity to identify novel disease genes as well as variants that confer disease-risk or modify a phenotype. Additionally, WGS data can be used to identify structural variation (SV) such as copy number variants (CNV) and chromosomal rearrangements (Liu et al., 2015; Timpson et al., 2018). As such, WGS is an attractive option for discovering novel genetic aberrations that cause or modify disease.

Until recently, the cost of WGS has been prohibitive. As such, the majority of

published studies have only reported WGS for individuals or small sample cohorts, which are often inadequate for family-based disease gene discovery, and certainly for association-based research. This is compounded by the fact that the amount of data produced by WGS is far greater than that from WES, meaning that computing power for analysis is also prohibitive. Though, as was seen with WES, as costs decline and data storage and analysis strategies evolve, the utility of WGS will continue to spread with an increase in accessibility (MacArthur et al., 2014).

The comprehensive set of genetic data made available from WGS will be required to solve many of the remaining genetic conditions. An example illustrating this is a study by Gilissen et al. (2014), who performed WGS of 50 patients with an intellectual disability and no genetic diagnosis after microarray and WES analysis. WGS identified 84 *de novo* single nucleotide coding variants and eight *de novo* copy number variants not detected previously, which with further analysis, led to genetic diagnoses in 20 of these patients. This highlights the value added by WGS and its promise to further our understanding of genetic conditions beyond what has been achieved with WES.

1.5.2.3 Targeted sequencing

Targeted sequencing is a customisable form of NGS that can be tailored to capture any region of the genome, whether coding, regulatory or intronic (Voelkerding et al., 2009). The genomic regions of interest can be captured by a pool of biotinylated RNA probes, microarrays, or PCR amplification, and subsequently undergo massively parallel sequencing (Liu et al., 2015). The massive reduction in genomic regions under examination reduces cost, time, storage and analysis requirements. It also allows far deeper coverage for each nucleotide of the targeted region, which in turn, minimises false positive and negative results (Fernandez-Marmiesse et al., 2018).

Most commonly, targeted sequencing has been used to screen for mutations in panels of known disease genes. This is an efficient way to screen genes in conditions with genetic heterogeneity or those caused by mutations in large genes that are difficult to PCR amplify and subsequently sequence. It is also useful for screening selected genes in large patient cohorts. The use of targeted sequencing rather than less biased WES or unbiased WGS, can reduce the number of incidental findings of mutations causing other diseases or those of uncertain significance (Liu et al., 2015). Further, the deep sequencing coverage afforded by targeted sequencing allows the detection of somatic and mosaic varinats, and is particularly useful in cancer (Fernandez-Marmiesse et al., 2018).

1.5.3 Twin studies

Twins have long been used as a tool for uncovering the genetic contribution to phenotypes. Monozygotic (MZ) twins share 100% of their DNA sequence and are genetically identical (with the exception of rare somatic or *de novo* germline mutations). In comparison, dizygotic (DZ) twins are genetically equivalent to any pair of siblings, sharing an average of 50% of their DNA sequence (Boomsma, 2013). Both MZ and DZ twins remove confounding factors such as age, pre- and (partially) post-natal environment differences. Twin-based heritability studies use these characteristics to estimate the contribution of genotype to phenotype. That is, phenotypic similarity between MZ twins and DZ twins is compared to assess environmental influences (differences between MZ twins) and genetic influences (similarity between MZ twins vs similarity between DZ twins) (Boomsma, 2013).

More recently, studies of MZ twins discordant for disease used WGS to identify *de novo* mutations that might be causing, or protecting against disease. Such studies identified aneuploidy discrepancies in MZ twins discordant for trisomy 13 (Ramsey et al., 2012), trisomy 21 (Dahoun et al., 2008) and X and Y chromosome aneuploidies (Razzaghian et al., 2010). CNV disparity has also been reported for MZ twins discordant for Parkinson's disease (Bruder et al., 2008) and congenital heart disease (Breckpot et al., 2012). Single nucleotide polymorphisms between discordant MZ twins are rare but have been observed in neurofibromatosis type 1 (Vogt et al., 2011).

1.6 Current state of ALS genetics research

Figure 1.2 provides a summary of the ALS genetic discoveries over the last 25 years. Linkage analysis facilitated the discovery of many ALS genes including *SOD1* (Rosen, 1993), *TARDBP* (Sreedharan et al., 2008), *FUS* (Vance et al., 2009), and *UBQLN2* (Deng et al., 2011), and was therefore a very powerful and successful tool for gene discovery. These discoveries were made using large ALS families, and as such, the causal gene mutations in most large ALS families have now been identified. More recently, NGS technologies have been instrumental in broadening the genetic spectrum of ALS with familial disease gene discoveries including *VCP* (Johnson et al., 2010), *PFN1* (Wu et al., 2012a), *HNRNP* genes (Kim et al., 2013), *MATR3* (Johnson et al.,

2014b), *CHCHD10* (Bannwarth et al., 2014), *CCNF* (Williams et al., 2016b) and *TIA1* (Mackenzie et al., 2017) and association of ALS with genes such as *TUBA4A* (Smith et al., 2014), *TBK1* (Cirulli et al., 2015; Freischmidt et al., 2015), *C21orf2* (van Rheenen et al., 2016) and *NEK1* (Kenna et al., 2016). However, there remain many genetic causes of ALS to be identified.

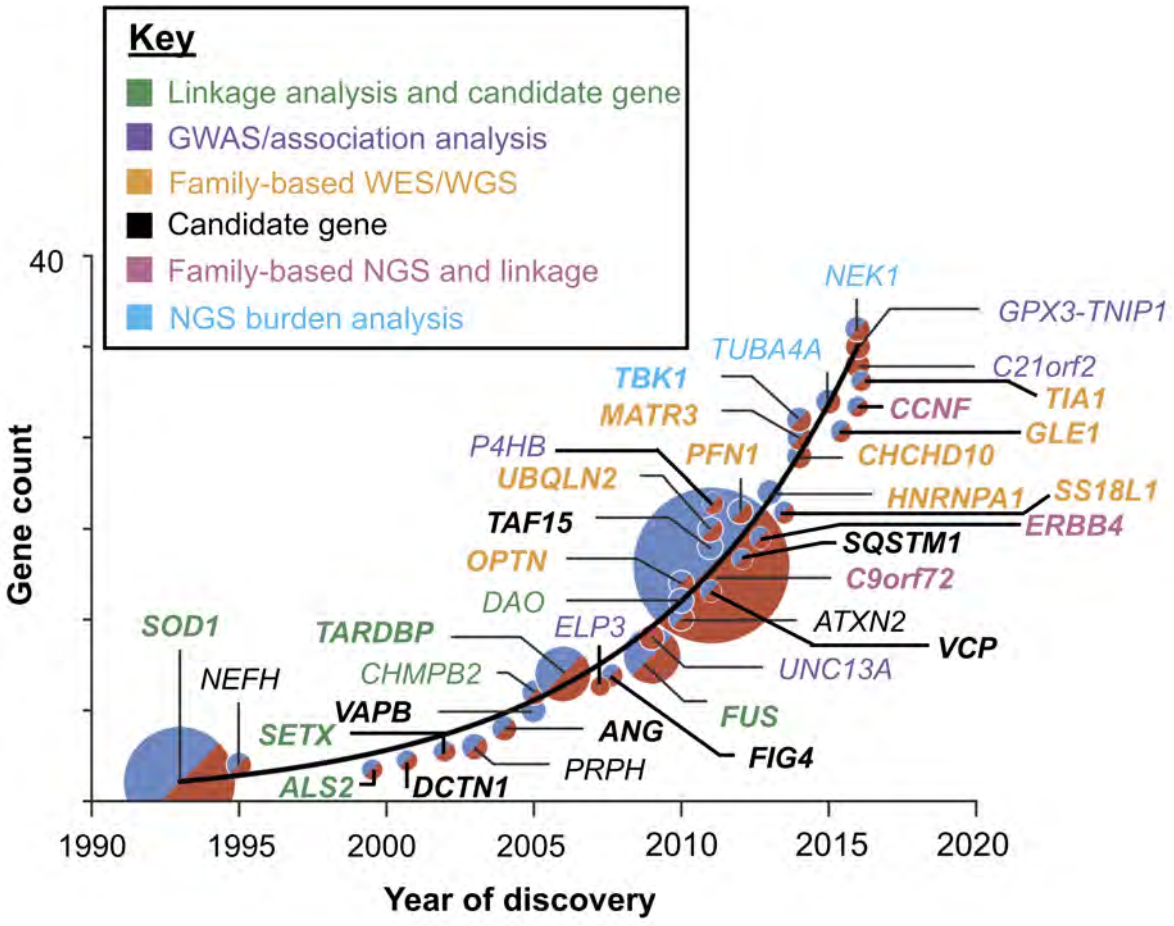


Figure 1.2: **Gene discovery in ALS over the last 25 years.** The number of reported ALS genes has grown drastically since the discovery of *SOD1* mutations in 1993. Circle size represents the proportion of familial ALS (FALS) patients who carry a mutation in that gene. Blue circles indicate genes linked only to FALS, red circles indicate those associated only with sporadic ALS (SALS), and circles half red and half blue represent genes implicated in both FALS and SALS. The colour of the gene name represents the methodology used for gene discovery. Genes in bold are those harbouring causal ALS mutations, while those in normal font have been associated with ALS, as shown in Table 1.1. Figure adapted from Brown and Al-Chalabi (2017).

ALS is genetically heterogeneous, with over 20 causal and many more associated

ALS genes identified over the past 25 years. From this, we know that genetic predisposition plays a major role in the development of ALS and is likely to contribute to those remaining cases with an unknown cause. However, this heterogeneity also suggests that the genetic factors causing disease in these remaining patients are likely complex, and will be challenging to uncover. Of the ALS patients with a family history of disease, one third carry an unknown causal gene mutation. These families represent our best opportunity to find genetic variants underlying disease, as their family history dictates that they possess a strong genetic predisposition to ALS.

1.6.1 Limiting factors for gene discovery in ALS

The remaining "unsolved" ALS families exhibit a number of characteristics that have hindered studies to identify their causal mutation. First and foremost, the majority of these families have reduced or incomplete penetrance of disease, where some mutation carriers do not develop ALS. Similarly, as a late onset disease, family members carrying ALS causal mutations may die from an unrelated event before they reach the age at which they would have developed ALS. This causes pedigrees to show an apparent skipping of generations where no one has developed disease, though their offspring do. This often means that DNA samples have not been collected from all informative family members. An obligate mutation carrier may never donate a DNA sample because there is no apparent need - they are not affected, and by the time disease develops in the proband, that parent has passed away, as has their affected parent (the probands grandparent), who appeared to be a sporadic patient when they presented with disease. Similarly, the variability in age of onset together with incomplete disease penetrance mean that we cannot assume unaffected family members (even those of 80-90 years of age) do not carry the causal mutation. As such, genetic studies are heavily reliant on DNA samples from affected individuals. This means that genetically speaking, many of these families are very small as there is often only one, two or three informative DNA samples available, causing both linkage and segregation analysis to be limited in power.

Heterogeneity is an additional barrier to effective gene discovery. The genetic heterogeneity of ALS means that linkage analysis in multiple families cannot be easily combined (as discussed in Section 1.5.1), which would otherwise be useful given the small nature of the remaining ALS families. There is also phenotypic heterogeneity including the overlap with FTD, and potentially other neurodegenerative conditions,

such as other forms of dementia. This too causes limited sample collection from families with multiple neurodegenerative conditions, as those individuals suffering from a related cognitive impairment rather than pure ALS may not be deemed informative. Familial classifications may also go unnoticed in such families.

The highly variable age of onset, incomplete penetrance and phenotypic heterogeneity in FALS patients have also led to confusion and discrepancy among clinicians when classifying familial and sporadic forms of disease, which can leave genetic research in a state of limbo. No clear consensus exists as to the exact criteria required to classify a patient as FALS (Al-Chalabi et al., 2017; Byrne et al., 2012a). Patients that have a first or second degree relative (a parent, sibling, grandparent, aunt, uncle, niece, nephew or half-sibling) who has also been diagnosed with ALS are invariably classified as FALS. However, in cases where the closest known relative also diagnosed with ALS is a third degree or higher, there is debate as to whether this constitutes FALS or whether the two are sporadic patients. As described in Section 1.3.2, the prevalence of ALS in Australia is just 8.7/100,000 people. As such, the probability of two related individuals within the same family both being affected by sporadic disease is exceptionally unlikely. Further, patients who have a relative affected by FTD should also be considered familial patients. Section 1.3.2 also described that ALS and FTD are considered to represent a spectrum of neurodegenerative disease with a shared genetic basis and pathology. As such, relatives with these conditions most likely have a common genetic mutation underlying disease.

The highly variable age of onset, incomplete penetrance and heterogeneity seen in ALS also suggests there may be other factors at play that are "switching on" disease onset. These could be epigenetic mechanisms impacting disease gene expression such as a particular DNA methylation pattern or acetylation on a certain histone. The possibility of this sort of unknown modifier influencing disease onset complicates disease gene discovery, as these remain unidentified, and therefore cannot be accounted for in our search for pathogenic gene mutations.

1.6.2 Novel strategies for gene discovery in ALS

While genetic linkage paved the way to the first wave of ALS gene discovery, and NGS the second wave, we have now reached another pivotal point in ALS gene discovery. Given the complex nature of the remaining ALS families, conventional techniques used

in isolation are inadequate for identifying the remaining ALS genes. To circumvent the challenges faced using these families, innovative strategies using multiple tools and approaches will be required.

It is becoming increasingly popular to use linkage analysis in combination with NGS to streamline the gene discovery process (Gazal et al., 2016; Ott et al., 2015). In days gone by, linkage analysis would identify a disease-linked locus, and a time-consuming candidate gene approach using Sanger sequencing would ensue. Today, NGS data can be used in place of Sanger sequencing to rapidly interrogate all candidate genes falling within the disease-linked region. The complementary approach would be that candidate causal mutations identified by NGS found to fall outside of disease-linked regions may be excluded. This approach was successfully applied by our laboratory in the discovery of the ALS gene *CCNF* (Williams et al., 2016b). Similarly, GWAS was used in combination with linkage analysis to identify the chromosome 9 linked ALS locus, which was subsequently found to harbour the most frequent known cause of ALS, hexanucleotide repeat expansions in *C9orf72* (DeJesus-Hernandez et al., 2011; Renton et al., 2011).

While NGS data can also be used for linkage analysis (Gazal et al., 2016; Ott et al., 2015), cost is a prohibitive factor, as only a few individuals may be sequenced. Thus linkage analysis using NGS data may be underpowered to detect significantly disease-linked regions. However, this approach can still be very useful for excluding unlinked genomic regions (Gazal et al., 2016; Ott et al., 2015). The Pedigree Variant Annotation, Analysis and Search Tool (pVAAST) program has been developed to facilitate the use of WES or WGS data in a linkage based model (Hu et al., 2014). Linkage analysis is combined with case-control association data and functional predictions that form the basis of its predecessor, VAAST (Hu et al., 2014; Yandell et al., 2011), in order to prioritise variants or genes that are associated with disease.

1.7 Project Aims

Gene mutations remain the only proven cause of ALS, some 25 years after the first ALS gene was identified. During this time, more than 20 new ALS genes have been uncovered. These genetic discoveries have provided the targets for numerous downstream research efforts into understanding the pathogenesis of ALS. For example, animal and cellular models of disease have been developed by introducing these ALS

causal mutations; molecular pathways disrupted during disease have been identified; and numerous mechanisms of disease have been proposed. Without the preceding genetic discoveries, none of these pivotal insights into ALS would have been possible.

These genetic discoveries have not only aided research but have also had a direct impact on patients and their families. ALS mutations have added to the diagnostic regimes available for patients, and the pre-symptomatic tests offered to their family members. Importantly, gene discoveries have also given families the opportunity to undergo preimplantation genetic diagnosis to prevent future generations from developing ALS. There are also numerous clinical trials, including gene therapies, underway which target molecules and pathways implicated in disease either directly or indirectly by genetic discoveries.

However, almost 90% of patients, including one third of FALS patients, have an unidentified underlying cause of ALS. The genetic heterogeneity of ALS suggests there are still many more genetic causes of disease to be uncovered. Studying families affected by ALS offers the clearest path towards further genetic discoveries, from which the findings can be extended to larger patient cohorts. The increasing utility of NGS technologies offers an exciting opportunity for the identification on novel genetic causes of ALS.

The aim of this project is to identify novel genetic mutations that cause ALS in Australian families with a history of ALS and to extend these findings to study patients with apparently sporadic ALS.

Specifically, the aims of this thesis are to:

1. Develop pipelines for handling large cohorts of next-generation sequencing data for gene discovery in ALS.
2. Investigate known and candidate ALS genes in familial and sporadic ALS patients to identify novel ALS mutations and/or associated genetic variants. (Paper I, Manuscript II)
3. Identify novel ALS genes and mutations in families with a history of ALS and in monozygotic twins discordant for disease. (Manuscript III)

"The best way to get started is to quit talking and begin doing."

Walt Disney

2

General subjects and methods

This Chapter describes the methods that form the foundation of data generation in this project. Next-generation sequencing (NGS) data was analysed throughout all Chapters of this thesis, and was the main tool used for gene discovery. Both whole exome (WES) and whole genome (WGS) sequencing have been utilised. WES was applied to FALS patients for known ALS gene (Chapter 4), candidate ALS gene (Chapter 5) and family-based gene discovery (Chapter 6) analyses. WGS data has been utilised for analysis of known (Chapter 4) and candidate (Chapter 5) ALS genes in SALS patients, family-based gene discovery in a medium-sized ALS family (Chapter 6) and analysis of ALS-discordant monozygotic twins (Chapter 7). SNP microarray genotyping was utilised for genetic linkage analysis of a medium-sized ALS family (Chapter 6), as well as for validation of selected WGS variants generated for ALS-discordant monozygotic twins (Chapter 7). A range of other genetic strategies were also employed to validate and evaluate the results obtained from the analysis of NGS data, including Sanger sequencing, segregation analysis, control genotyping and *in silico* assessment tools.

2.1 Subjects and patient cohorts

2.1.1 Patient recruitment and sample collection

The majority of ALS patient samples were collected from two clinics; the Macquarie Neurology Clinic, directed by Professor Dominic Rowe, and The Molecular Medicine Laboratory, Concord Hospital, directed by Professor Garth Nicholson. Patients of the Macquarie Neurology clinic and their family members were recruited to the Macquarie University Neurodegenerative Disease Biobank for research participation. Additional samples were also collected from the Australian MND DNA bank, Royal Prince Alfred Hospital. The vast majority of patients were of European descent, and all patients were diagnosed with probable or definite ALS according to El Escorial criteria (Brooks et al., 2000). DNA was extracted from peripheral blood using standard protocols. Manual protocols were applied to those samples obtained from the Molecular Medicine Laboratory and the Australian MND DNA bank, while the QIASymphony automated liquid handling robot and the DSP Midi extraction kit (Qiagen) were utilised for samples collected from the Macquarie University Neurodegenerative Disease Biobank.

2.1.2 Ethics and consent

Each patient (affected individual) and control individual provided informed written consent to be involved in genetic research as set out by the Human Research Ethics Committees of Macquarie University (Approval number 5201600387) and the Sydney South West Area Health Service (Approval number CH62/6/2011-123-G Nicholson HREC/11/CRGH/179, Title: Research study into identifying new gene mutations for motor neuron disease).

2.1.3 Patient cohorts

Various patient cohorts underwent genetic analysis as part of this thesis. A summary of these cohorts, and the sequencing approaches applied to each, is provided in Figure 2.1. Some cohorts were only used as part of the development of bioinformatic processing scripts and pipelines in Chapter 3, and were not directly utilised for genetic discovery in this thesis.

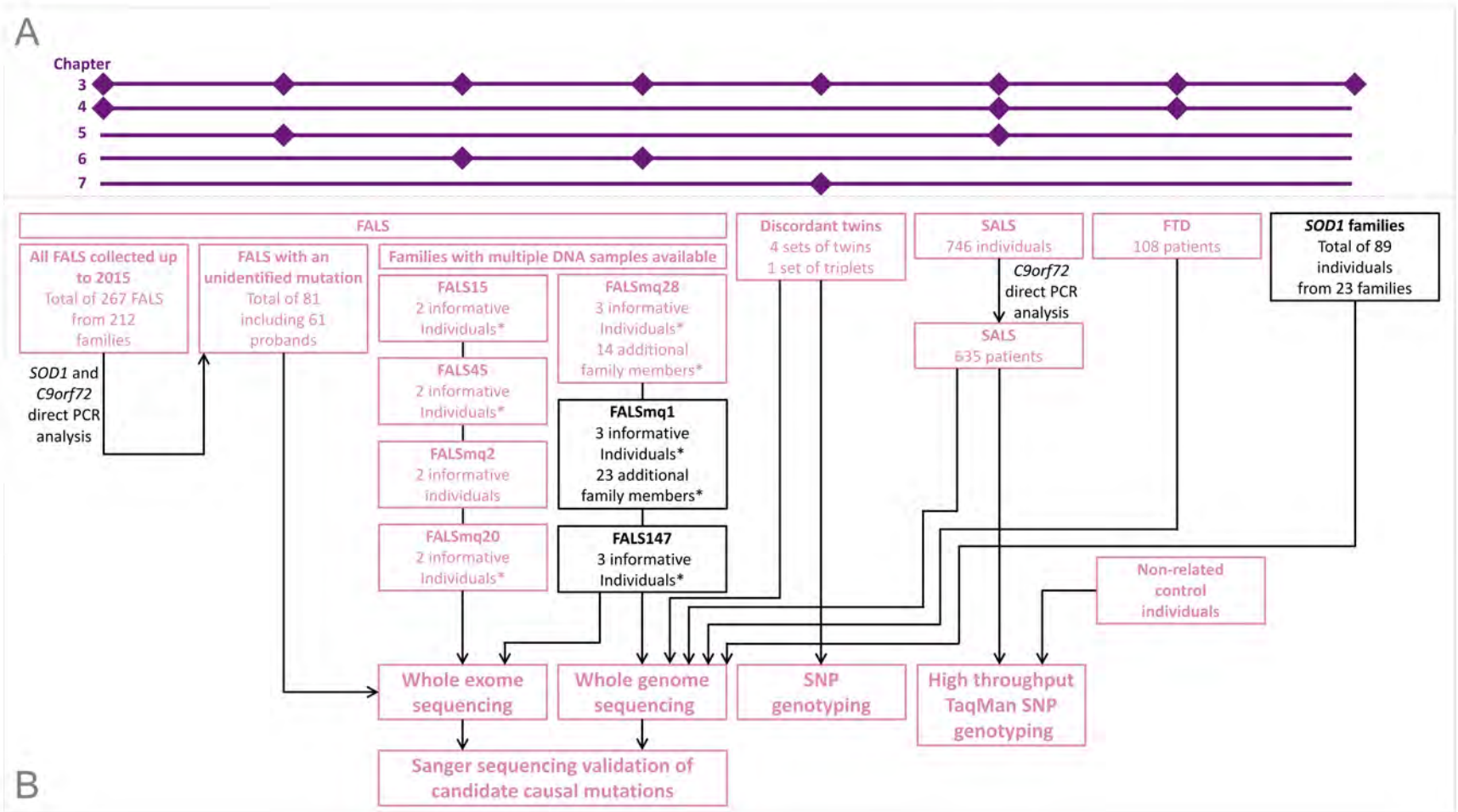


Figure 2.1: **Patient cohorts and sequencing datasets.** (A) Cohorts used in each Chapter of this thesis. Purple lines are used to indicate the Chapters in which each cohort was used, with diamonds representing the cohorts used in the major genetic analysis component of a given Chapter. (B) Genetic sequencing techniques applied to ALS patient cohorts. Pink boxes indicate the patient cohorts directly utilised in the discovery analyses presented in this thesis (Chapters 4 – 7), while black boxes indicate those which were incidentally analysed as part of some bioinformatics processing pipelines developed in Chapter 3. The arrows indicate the genetic sequencing technique(s) applied to each cohort. *Informative family members were those that were either ALS patients, obligate mutation carriers or the "married-in" control parent of an ALS patient. Additional family members were those that were "at-risk" of carrying an ALS causal mutation, or the "married-in parent" of an "at-risk" individual. Additional family members did not undergo WES or WGS and were only used for SNP microarray genotyping.

2.1.3.1 Familial ALS (FALS)

Each family member of a FALS family was classified into one of the four subject groups described below. Among these subject types, affected individual/patient, obligate mutation carrier and "married-in" control individuals were considered informative for family-based genetic analysis. "At-risk" individuals were considered additional family members.

Affected individual/patient: Diagnosed with ALS. Further, a proband patient was the first member of their family to present at one of the above clinics, and in many cases was the only member of their family for whom a DNA sample was available.

Obligate mutation carrier: An individual who must carry the causal gene mutation. This refers to an unaffected individual who has both a direct ancestor and direct descendent affected by ALS, and therefore must have inherited, and passed on the causal gene mutation. This individual may go on to develop ALS later in life, or may die without ever developing disease. This may be due to the variable age of disease onset with death due to another cause prior to reaching the age at which they would have experienced disease onset, or reduced penetrance of the causal gene mutation.

"Married-in" control: The spouse of a patient or obligate mutation carrier, who is considered to be unrelated to their partner, as determined by family history. In many cases, these individuals are also the parent of an ALS patient or "at-risk" individual.

"At-risk" individual: An individual who is unaffected at the time of recruitment but is "at-risk" of developing ALS (and carrying the causal gene mutation), based on their descent from an affected individual. These individuals do not have any affected direct descendants (though these descendants also possibly carry the causal mutation and may go on to develop disease later in life), and therefore there is no way to be certain whether they carry the causal mutation or not.

2.2 Next generation sequencing (NGS)

As described in Chapter 1, Section 1.5.2, NGS is based on massively parallel sequencing methods used to generate a vast volume of DNA sequence information. Table 2.1 outlines the methods used to generate WES and WGS data using Illumina sequencing based protocols, for the various cohorts utilised in this thesis. Generation of both WES and WGS data employed the basic steps of library capture, cluster generation, sequencing and data processing, as summarised in Figure 2.2, and described in the following sections. WES data, other than that for family FALSmq28, was generated prior to commencement of candidature, while all WGS data and FALSmq28 WES data was generated during candidature.

2.2.1 Generation of raw sequencing data

NGS was performed by sequencing providers, as described in Table 2.1. The general principles underlying both WES and WGS were largely similar (Figure 2.2). The major difference was the initial library preparation phase, as shown in Figure 2.2A. The first step for both WES and WGS was the shearing of genomic DNA (gDNA) by either mechanical sonication or biological enzymatic digestion, in order to produce DNA fragments (Metzker, 2010; van Dijk et al., 2014), which were then ligated to adapter oligonucleotides (Seaby et al., 2016; Zhang, 2014). At this point during WES, coding sequences were enriched by hybridisation to exome-complementary probes (Bamshad et al., 2011; Teer and Mullikin, 2010), while for WGS, bead-based size selection was applied to enrich for fragments 150bp in length. Following this fragment selection, PCR amplification was performed for WES, as well as in special cases of WGS where gDNA availability was low.

As depicted in Figure 2.2B, the Illumina (California, United States) bridge-PCR amplification approach (Illumina, 2017) was employed to generate DNA fragment clusters, to ensure sufficient sequencing signal was generated. The key component of this cluster generation was the flow cell, a solid matrix covered in forward and reverse primers complementary to the adapter sequences added during library preparation. Upon application of the DNA library to this flow cell, one end of an adapter ligated DNA fragment bound a primer, while the remaining free end paired with the immediately adjacent opposite primer, in order to adopt a bridge formation (Casals et al.,

TABLE 2.1: **Details of NGS data generation.**

Cohort(s)	No. samples	Genetic data type	Service provider	Library preparation kit	Coverage		Average read depth	Sequencing platform	Quality control	Alignment	Joint calling	Variant calling	Annotation*	Output file	Genetic discovery Chapter
					Exons	Mb				BWA version	GATK version	GATK version			
FALS	113	WES	Macrogen	Illumina TruSeq	201,121	64	100X	Illumina HiSeq2000	FastQC	v0.7.12	v3.4.0	v3.4.0	ANNOVAR	137-sample VCF	4, 5, 6
FALS	24	WES	Macrogen	AgilentSureSelect AllExon+UTRV5	359,555	75	100X	Illumina HiSeq2000	FastQC	v0.7.12	v3.4.0	v3.4.0	ANNOVAR	137-sample VCF	4, 5, 6
FALSmq28	3	WES	Macrogen	AgilentSureSelect AllExon+UTRV5	359,555	75	100X	Illumina HiSeq4000	FastQC	v0.7.12	N/A	v3.4.0	ANNOVAR	individual VCF	6
	3	WGS	Macrogen	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	N/A	v3.7	ANNOVAR	individual VCF	6
	3	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	6
FALSmq1	3	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.
FALS147	3	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.
SALS	628	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.
FTD	108	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.
SOD1 families	89	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.
SALS female twins∧	2	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	7
SALS male twins∧	2	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	7
SOD1 female triplets∧	3	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	7
C9orf72 male twins∧	2	WGS	Kinghorn	Illumina TruSeq DNA Nano	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	7
Duplicate samples	7	WGS	Kinghorn	Illumina TruSeq DNA PCR-Free	N/A	3,000	30X	Illumina Hiseq X Ten	FastQC	v0.7.15	v3.7	v3.7	ANNOVAR	850-sample VCF	.

Abbreviations: BWA, Burrows Wheeler Aligner; and GATK, Genome Analysis ToolKit.

*Annotation using ANNOVAR software was performed by the candidate for all cohorts except A and B.

∧Twin data also underwent alignment and variant calling using Isaac software as described in Chapter 7.

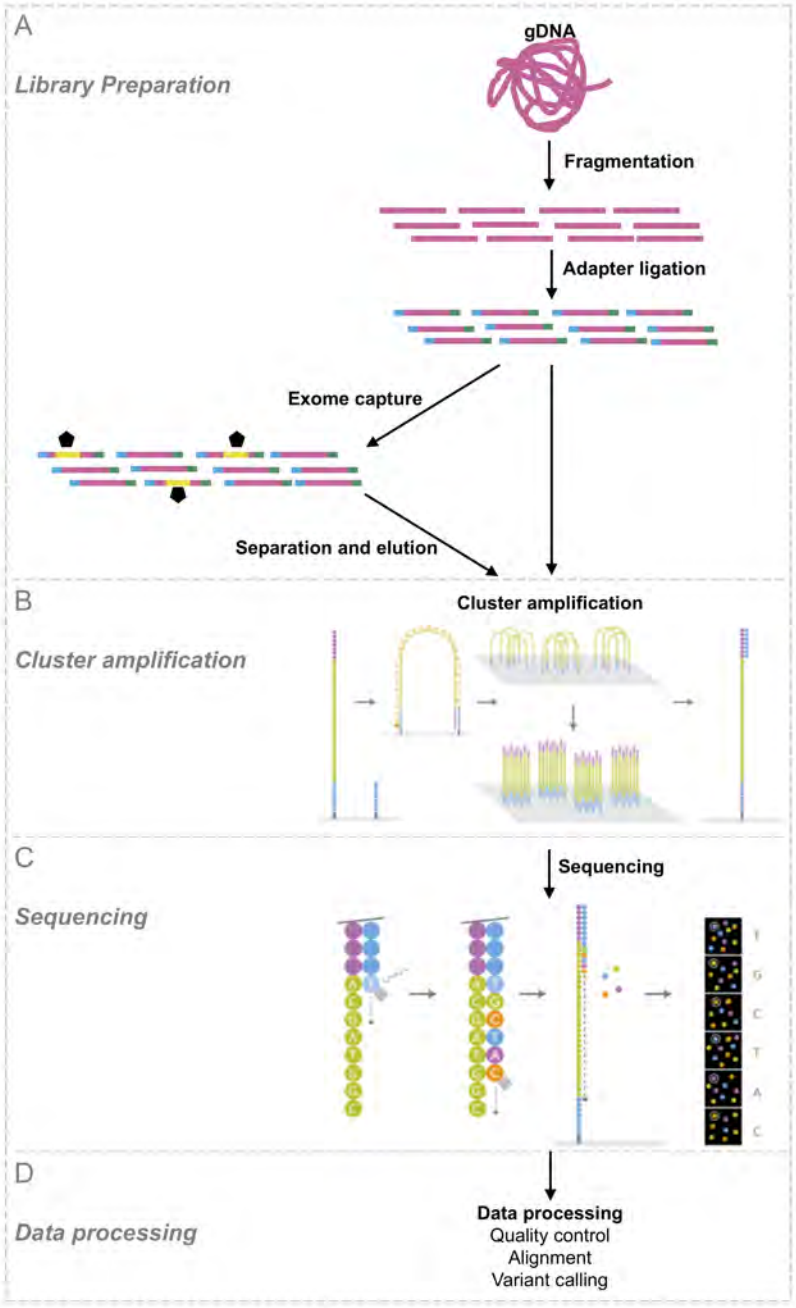


Figure 2.2: **General Illumina sequencing work flow.** (A) Library preparation. DNA was fragmented and ligated with adapter oligonucleotides, and in the case of WES, enriched for exonic sequences. (B) Cluster amplification. The DNA library was loaded on and hybridised to the flow cell. Each bound DNA fragment was then amplified to a clonal cluster through bridge-PCR amplification. (C) Sequencing. Following addition of sequencing reagents, the sequencing primers bound the adapter sequence to facilitate sequencing. Fluorescent nucleotides were incorporated into the amplification product, causing a corresponding light emission, which was then imaged. The emission signal from each cluster was used to determine the identity of the DNA base. The cycle was then repeated up to 150 times to produce sequence reads of 150bp. (D) Data processing. Raw sequencing reads then underwent bioinformatic processing for quality control, alignment to the reference sequence and variant calling.

2012). Each such fragment then underwent PCR amplification to produce distinct clonal clusters, leaving the fragments ready for sequencing.

The gold standard Illumina sequencing-by-synthesis chemistry was then applied, as shown in Figure 2.2C. This involved a proprietary reversible terminator-based method, which detects single nucleotide bases as they are incorporated as part of the extension product (Illumina, 2017). Tens of millions clonal clusters were sequenced in parallel using this method.

2.2.2 Data processing

2.2.2.1 Quality control, alignment and variant calling

Standard data processing was also performed by sequencing service providers as described in Table 2.1. Quality control using proprietary Illumina methodology and the FastQC program (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was applied to remove or trim low quality raw sequencing reads (FASTQ format) prior to further processing. Alignment, or mapping, of raw sequencing data to the reference genome, GRch37 (hg19), was then completed using the Burrows Wheeler Aligner (BWA) (Li and Durbin, 2009, 2010) or Isaac Aligner (Raczy et al., 2013). This produced SAM (Sequence Alignment/Map) files, which were converted to a binary format (BAM), using SAMtools software (Li et al., 2009). These alignment files were then used to call variants using either the Genome Analysis ToolKit (GATK; McKenna et al., 2010) or Isaac (Raczy et al., 2013) variant callers. This processing pipeline produced a variant call file (VCF) for each sample. These VCFs contained variant specific information including chromosomal location, reference and alternate nucleotides, alongside quality measurements and genotype data (details provided in Chapter 3, Section 3.2). Single nucleotide polymorphism (SNP) and small insertion/deletion (indel) variant calls were ultimately used to assess genetic variation differences between affected individuals and controls.

2.2.2.2 Variant annotation

Annotation of NGS variant data with biologically meaningful information is imperative for downstream filtering and interpretation. A number of software programs, including ANNOVAR (Wang et al., 2010, http://annovar.openbioinformatics.org/en/latest/) have been developed to integrate information from multiple biological *in*

silico databases with VCFs. WES (other than that generated for family FALSmq28) data was annotated by Dr Qiongyi Zhao (University of Queensland), whereas all other NGS data was annotated by the candidate (using Script 2.1). A summary of the databases utilised in the ANNOVAR annotation performed by the candidate is provided in Table 2.2. All variants were annotated with the gene in which each variant resided within (or was in closest proximity to) and were also assigned to a genomic functional category, being one of exonic, splicing, ncRNA, UTR5, UTR3, intronic, upstream, downstream or intergenic. All exonic variants were further classified as a frameshift insertion, frameshift deletion, frameshift block substitution, stopgain, stoploss, non-frameshift insertion, non-frameshift deletion, non-frameshift block substitution, non-synonymous SNV (single nucleotide variant), synonymous SNV or unknown. The dbNSFP (Database for Non-Synonymous SNPs' Functional Predictions; Liu et al., 2011) was utilised to add the predicted functional effect of each variant, from various protein prediction programs, to the VCF. Additional information pertaining to the absence/presence and/or minor allele frequency (MAF) of each variant in control databases (including dbSNP, ExAC and gnomAD; see Table 2.4 for details) was also used in annotation.

Code 2.1: **ANNOVAR.sh** This script was used to annotate a VCF) with information from the databases listed in Table 2.2, using the ANNOVAR software tool.

```
#!/bin/sh
#
#  ANNOVAR.sh

# perform ANNOVAR annotation for EXAMPLE.vcf
perl table_annovar.pl EXAMPLE.vcf humandb/ -buildver hg19 -out myanno
    -remove -protocol refGene,cytoBand,exac03,gnomad_exome,gnomad_genome,
    avsnp147,dbnsfp33a,dbnsfp31a_interpro,esp6500siv2_ea,esp6500siv2_all,
    ALL.sites.2015_08,EUR.sites.2015_08,clinvar_20170130 -operation
    g,r,f,f,f,f,f,f,f,f,f,f -nastring . -vcfinput
```

Table 2.2: **Databases used for ANNOVAR annotation.**

Annotation	Description
refGene	Closest gene according to RefSeq Gene
cytoBand	Giemsa-stained chromosomes bands
exac03	ExAC exomes allele frquency data - from all populations and each individual population including Non-Finnish Europeans
gnomad_exome	gnomAD exomes allele frquency data - from all populations and each individual population including Non-Finnish Europeans
gnomad_genome	gnomAD genomes allele frquency data - from all populations and each individual population including Non-Finnish Europeans
avsnp147	dbSNP147 membership/ID
dbnsfp33a	Protein prediction scores from dbNSFP v3.3a
dbnsfp31a_interpro	Protein domain
esp6500siv2_ea	NHLBI-ESP project allele frequency data from European American populations
esp6500siv2_all	NHLBI-ESP project allele frequency data from all populations
ALL.sites.2015_08	1000Genomes allele frequency data from all populations
EUR.sites.2015_08	1000Genomes allele frequency data from European populations
clinvar_20170130	Clinvar classifications

Abbreviations: ExAC, Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; dbNSFP, Database for Non-Synonymous SNPs' Functional Predictions; and NHLBI-ESP, National Heart, Lung, and Blood Institute - Exome Sequencing Project.

2.3 Genome-wide SNP microarray genotyping

SNP microarrays are used to genotype hundreds of thousands of common SNPs in parallel, and are also amenable to high-throughput use with large sample cohorts. In this project, the Infinium CoreExome-24 BeadChip v1.0 and v1.1 (Illumina) microarrays were used to genotype a total of 547,644 and 551,839 SNP markers, respectively. Raw data was generated and processed by service providers as described in Table 2.3. For each SNP marker, a 50bp oligonucleotide probe complementary to the site adjacent sequence was immobilised on a solid surface. During the reaction, fragmented DNA bound the probe, and underwent single-base extension with a fluorescently labelled nucleotide complementary to the SNP, which caused emission of the corresponding intensity signal. This signal intensity data was recorded by an iScan system (Illumina), and was subsequently processed using the GenomeStudio v2011 (Illumina) genotyping module to genotype each SNP marker independently.

Table 2.3: **Details of SNP microarray genotyping data generation.**

Cohort(s)	**No. samples**	**Platfrom**	**Scanner**	**No. SNP markers**	**Genotype calling**	**Marker & pedigree cleaning**	**SNP pruning**	**Output file**	**Genetic discovery Chapter**
FALSmq28	16	Illumina InfiniumCoreExome-24v1-1	Illumina iScan	551,839	GenomeStudio v2011	PedStats	PLINK v1.07	.ped with genotype data; and ancillary .map and .dat files	6
SALS female twins	2	Illumina InfiniumCoreExome-24v1-1	Illumina iScan	551,839	GenomeStudio v2011	N/A	N/A	.idat	7
SALS male twins	2	Illumina InfiniumCoreExome-24v1-1	Illumina iScan	551,839	GenomeStudio v2011	N/A	N/A	.idat	7
SOD1 female triplets	3	Illumina HumanCoreExome-24v1-0	Illumina iScan	547,644	GenomeStudio v2011	N/A	N/A	.idat	7
C9orf72 male twins	2	Illumina HumanCoreExome-24v1-0	Illumina iScan	547,644	GenomeStudio v2011	N/A	N/A	.idat	7

2.4 NGS variant validation strategies

Each genomic variant identified from genetic analysis of NGS data with the potential to cause disease either within an ALS family, discordant monozygotic twin pair or single affected individual (from candidate gene screening), was considered a candidate mutation (eg. novel, non-synonymous variants). Even following quality control filtering, there remains the potential for any variant identified by NGS to be a sequencing artefact. Thus the true presence of each candidate mutation within the relevant DNA sample needed to be confirmed. This was achieved using NGS read visualisation and Sanger sequencing. Further, to determine the potential of the candidate mutation to cause disease, its novelty needed to be established by comparison to extensive numbers of healthy control individuals. Finally, in order to assess the potential pathogenic nature of a candidate mutation, a variety of additional *in silico* analyses were employed.

2.4.1 NGS read visualisation with the integrative genomics viewer

As described above in Section 2.2.2.1, NGS reads were aligned to the reference sequence during standard bioinformatics processing to facilitate variant calling. The integrative genomics viewer (IGV, Robinson et al., 2011) was used to visualise all aligned reads at any given position in the genome, using BAM files. Each candidate mutation was visually analysed in this way to determine whether there were sufficient high quality reads to support the variant call. Generally, at least 20% of all reads at the given position were required to possess the alternate allele to support a heterozygous variant call, and the majority of these were required to fall within the middle 90% of the sequence read (ie. not at the 5' or 3' extremity of the read).

2.4.2 Sanger sequencing of candidate mutations and segregation analysis

Massively parallel sequencing has higher error rates than Sanger sequencing (Pabinger et al., 2014; Zhang, 2014), and therefore variant validation by Sanger sequencing remains the gold standard. As such, all candidate mutations underwent Sanger sequencing, including PCR amplification and Sanger sequencing. Those found to be absent from affected individuals were false positives, and discarded from analysis. Similarly, candidate mutations identified in any "married-in" family control individuals

by Sanger sequencing were also false positive candidate mutations, and discarded as potentially pathogenic. Sanger sequencing of additional family members whom did not undergo WES or WGS also allowed the establishment of co-inheritance of candidate mutations within ALS family pedigrees, further establishing a genotype-phenotype relationship.

Gene or variant specific primers were designed using either ExonPrimer (http://ihg.gsf.de/ihg/ExonPrimer.html) or Primer3 Plus (http://bioinfo.ut.ee/primer3-0.4.0/), and were synthesised by Sigma Aldrich (NSW, Australia). MyTaq HS Red Mix (Bioline, London, United Kingdom) was used in all PCR reactions, and 10X PCR enhancer (Life Technologies, CA, USA) was added when required. In cases where variants were found within especially repetitive or duplicated genomic regions, touchdown thermocycling and/or nested PCRs were used. Further, some indel variants were validated using fragment length analysis of fluorescently labelled PCR products. Primer sequences, and optimised conditions are provided in Appendix A.3, Table A.1.

2.4.3 Control genotyping

To determine whether candidate mutations were in fact rare population specific variants rather than potential pathogenic mutations, or to establish the population frequency of potentially disease-associated variants, screening of large numbers of non-related, age- and population-matched control individuals was required.

2.4.3.1 Control database screening

The publicly available NGS control databases listed in Table 2.4 were inspected through the web browser interface, or by interrogation of VCFs using custom bioinformatics pipelines developed in Chapter 3, Section 3.5.3. Where appropriate, when analysing the ExAC and gnomAD databases, the Non-Finnish European (NFE) subset of individuals was primarily utilised due to the absence of an Australian European subset of controls in these databases.

TABLE 2.4: **Control databases used in this project.**

Control database	Total number of individuals	ancestry	Total number of variants	VCF file size	Institute	Literature reference	Web browser address
ExAC	60,706	Variable	9,362,538	34.1GB	Broad Institute	(Lek et al., 2016)	http://exac.broadinstitute.org/
gnomAD exomes	123,136	Variable	15,014,744	69.64GB	Broad Institute	(Lek et al., 2016)	http://gnomad.broadinstitute.org/
gnomAD genomes	15,496	Variable	4,500,726	19.42GB	Broad Institute	(Lek et al., 2016)	http://gnomad.broadinstitute.org/
MGRB	1,144	Australian	39,283,402	1.25TB		none	https://sgc.garvan.org.au
DACC	967	Australian	1,630,808	1.59GB	Diamantina Institute	none	N/A

Abbreviations: ExAC, Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; MGRB, Medical Genome Reference Bank; DACC, Diamantina Australian Control Collection.

2.4.3.2 TaqMan control genotyping

Custom TaqMan genotyping assays (Life Technologies) were designed and applied to Australian ALS affected individuals and control individuals to determine the frequency of potentially disease-associated variants. The TaqMan genotyping assay is based around the distinct fluorescent labelling of two allele specific probes, which are used in conjunction with a primer pair to amplify and label each allele at a particular genomic site. Standard thermocycling was performed on the ViiA7 RealTime System (Life Technologies), which also measured the fluorescence signals generated by each sample. Viia7 software then processed these signals to plot the fluorescence value for each sample, and evaluate whether a homozygous wild-type, heterozygote or homozygous variant genotype was present.

2.5 *In silico* tools and databases for assessment

In silico assessment of genetic variants can provide various lines of evidence to either support or refute their potential for pathogenicity. These insights can aid in determining which candidate mutations have the highest potential to cause ALS, and therefore warrant further investigation by *in vitro* or *in vivo* analyses. Table 2.5 describes each *in silico* tool and database utilised in this thesis for assessing the potential pathogenicity of candidate mutations. The following sections outline how each of these tools were implemented. In Chapter 6, these tools were used in combination to develop a pipeline and scoring system to assess potential ALS pathogenicity. This pipeline and scoring system was validated using known ALS gene mutations, and was subsequently

utilised to prioritise which candidate mutations were most likely to cause disease based on the similarities they showed with known ALS causal mutations. The guidelines for interpreting sequence variants published by the American College of Medical Genetics and Genomics (ACMG; Richards et al., 2015), as shown in Appendix A.3.3, were consulted as part of the development of this pipeline. As such, the rationale for the relevance of each characteristic to potential pathogenicity is discussed in Chapter 6.

Protein predictions

Protein prediction programs provide information about the likely structural and functional effect of genetic variation on the encoded protein. As part of this thesis, eight protein prediction programs were utilised (see Table 2.5 for details). Each program uses a complex algorithm based on structure and/or conservation to predict the potential effect of a genetic variant.

Species conservation analysis

The conservation of the amino acid affected by each candidate mutation was analysed across multiple species using three approaches. Firstly, all available known protein sequences were obtained from HomoloGene (`http://www.ncbi.nlm.nih.gov/homologene`), and subsequently aligned using Clustal Omega v1.2.4 (`http://www.ebi.ac.uk/Tools/msa/clustalo`; Sievers et al., 2011). The amino acid of interest was then manually assessed for conservation by calculating the percentage match of the human residue with the homologous residue in other species. Lastly, the software packages PhyloP (Pollard et al., 2010) and PhastCons (Siepel et al., 2005) were accessed through the UCSC (University of California Santa Cruz) genome browser table utility (`https://genome.ucsc.edu`), to score the degree of conservation of each amino acid of interest.

Gene expression

The level of gene expression in the brain was assessed by analysing the gene-specific graph available from the Human Brain Transcriptome (HBT) website (`http://hbatlas.org/`; Kang et al., 2011; Pletikos et al., 2014), and recording the signal intensity in the cerebellar cortex at the 14th lifetime period (approx. 80 years of age). Gene expression levels in the spinal cord were assessed using the Genotype-Tissue Expression Project (GTex) database (`https://www.gtexportal.org/home/`; Carithers et al., 2015) and were determined by analysis of the gene expression plot of interest for the "Brain - Spinal cord (cervical c-1)" tissue type.

Genic tolerance

Tolerance for genetic variation was assessed for each gene of interest by analysis of two metrics. Firstly, the residual variation intolerance (RVIS) score and its associated percentile score (http://genic-intolerance.org/; Petrovski et al., 2013). This percentile score indicated the percentile of most intolerant human genes within which the gene of interest fell. Secondly, the ExAC (http://exac.broadinstitute.org/; Lek et al., 2016) z-score constraint metric for missense variants was utilised, which indicated the degree of deviation between the number of missense variants observed in the gene in the ExAC database (among healthy control individuals), compared to that which was expected based on the size of the gene.

Gene/protein description

A description of each gene of interest was obtained from the GeneCards website (http://www.genecards.org/) gene summary page.

Prior implications in neurodegenerative disease

The PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/) was queried for the gene (and protein) name and the term "neurodegenerative disease" as follows; "<gene name> AND neurodegenerative disease". The number of matching entries was recorded, as were any publications of interest.

Protein structure

The protein sequence of interest was obtained from the UCSC genome browser (https://genome.ucsc.edu) and analysed using the SMART web tool (Simple Modular Architecture Research Tool; http://smart.embl-heidelberg.de/; Letunic et al., 2015) to determine which protein domains were present within the protein, and to subsequently identify the protein domain in which the genetic variant of interest fell. To determine whether any post-translational phosphorylation sites had been added or removed by the genetic variant of interest, both the canonical and mutant protein sequences were analysed using NetPhos 2.0 (http://www.cbs.dtu.dk/services/NetPhos-2.0/; Blom et al., 1999). The resultant predicted phosphorylation sites were then compared to determine whether the variant introduced or removed any predicted phosphorylation sites.

TABLE 2.5: ***In silico* tools utilised to assess the potential pathogenicity of candidate mutations.**

Tool/Database	Name	Description	Scores/Output	Website	Reference
Protein Predictions					
MutationAssessor	Functional impact of protein mutations	Assesses functional impact using evolutionary conservation of the affected residue in protein homologs	Predicted functional (high, medium), predicted non-functional (low, neutral)	http://mutationassessor.org/r3/	Reva et al. (2011)
MutationTaster	Mutation Taster	Uses evolutionary conservation, splice-site changes, loss of protein features and mRNA expression modifying features to predict function effect	Disease causing or polymorphism	http://www.mutationtaster.org/	Schwarz et al. (2014)
Polyphen-2	Polymorphism Phenotyping v2	Uses species sequence homology to predict the effect of amino acid substitution on protein function	Probably or possibly damaging or benign	http://genetics.bwh.harvard.edu/pph2/	Adzhubei et al. (2010)
Pon-P2	Pathogenic-or-Not-Pipeline	Predicts functional effect based on amino acid features, Gene Ontology (GO) annotations and evolutionary conservation	Pathogenic, neutral or unknown tolerance	http://structure.bmc.lu.se/PON-P2/	Niroula and Vihinen (2015)
SIFT	Sorting Intolerant From Tolerant	Uses sequence alignment and degree of amino acid residue conservation between closely related sequences to predict functional consequence	Damaging or tolerated	http://sift.jcvi.org/	Kumar et al. (2009)
PROVEAN	Protein Variation Effect Analyzer	Employs a generalised approach to assess functional effect on a protein	Deleterious or neutral	http://provean.jcvi.org/index.php	Choi et al. (2012)
SNPs&GO	Predicting disease associated variations using GO terms	Utilises protein sequence, evolutionary, and functional information (according to GO terms) to make predictions	Disease or neutral	https://snps-and-go.biocomp.unibo.it/snps-and-go/	Calabrese et al. (2009)
CADD	Combined Annotation Dependent Depletion	Uses annotation information on conservation, functional genomics, transcript position, gene expression and protein scores	Magnitude of rank score (10=top 10% deleterious, 20=top 1% deleterious etc)	http://cadd.gs.washington.edu/info	Kircher et al. (2014)
Species conservation analysis					
NCBI homologene	National Centre for Biotechnology Information homologene tool	System for collectimng homology data for gene sets from eukaryotic species	Protein sequences from various species	http://www.ncbi.nlm.nih.gov/homologene	
ClustalOmega	Multiple sequence alignment	Generates alignments between three or more sequences	Text file containing aligned sequences with residue counts	http://www.ebi.ac.uk/Tools/msa/clustalo	Sievers et al. (2011)
PhyloP	Phylogenetic Model	Uses a model of neutral evolution and alignment strategies to calculate conservation or acceleration p-values	Positive (conserved) or negative (accelerated)	http://compgen.cshl.edu/phast/index.php	Pollard et al. (2010)
PhastCons	Phylogenetic Analysis with Space/Time models - Conservation	Computes the probability of a nucleotide belonging to a conserved element	Between 0-1 with 1 being conserved	http://compgen.cshl.edu/phast/index.php	Siepel et al. (2005)
Gene expression analysis					
HBT	The Human Brain Transcriptome	Public database containing transcriptome data from the developing and adult human brain	An expressed gene is defined as one for which expression levels are greater than six on the log-2 signal intensity scale	http://hbatlas.org/	Kang et al. (2011); Pletikos et al. (2014)
GTex Project	Genotype-Tissue Expression Project	Web resource with data for gene expression, regulation and its relationship to genetic variation	Reads per kilobase of transcript per million (RPKM) values	https://www.gtexportal.org/home/	Carithers et al. (2015)
Genic tolerance					
RVIS	Residual Variation Intolerance Score	Public database with scores describing tolerance to genetic variation affecting gene and/or protein function	Percentage indicating the rank of intolerance (ie. 10%, top 10% of most intolerant genes)	http://genic-intolerance.org/	Petrovski et al. (2013)
ExAC	Missense contraint metric	Constraint metric describing the deviation from the expected number of missense variants in a gene calculated using variants found in ExAC control database	Z-score, positive (intolerant to variation) or negative (tolerant to variation)	http://exac.broadinstitute.org/	Lek et al. (2016)
Gene/protein description					
GeneCards	Entrez Gene Summary	Summary of the basic behaviour, pathway involvement and/or localisation of the encoded protein	Descriptive text	http://www.genecards.org/	N/A
Prior implications in neurodegenerative disease					
PubMed	PubMed	Database for biomedical literature		https://www.ncbi.nlm.nih.gov/pubmed/	N/A
Protein structure assessments					
SMART	Simple Modular Architecture Research Tool	Web resource providing identification and annotation of protein domains, based on manual curation from UniProt, Ensembl and STRING	Graphical and tabular representation of protein domains and the residues involved in each	http://smart.embl-heidelberg.de/	Letunic et al. (2015)
NetPhos 2.0	NetPhos 2.0	Predicts protein phosphorylation sites based on sequence and structure information	Text file indicating residues predicted to be phosphorylated	http://www.cbs.dtu.dk/services/NetPhos-2.0/	Blom et al. (1999)
Interacting partners					
STRING	STRING	Database of known and predicted protein protein interactions	Graphical and tabular representation of protein interactors for the given protein	http://string-db.org/	Szklarczyk et al. (2015)
BioGrid	Biological General Repository for Interaction Datasets	Curated repository of physical and genetic interactions	Tabular representation of protein interactors for the given protein	http://thebiogrid.org/	Chatr-Aryamontri et al. (2015)

Abbreviations: mRNA, messenger RNA; GO, Gene Ontology; ExAC, Exome Aggregation Consortium; and RPKM, Reads per kilobase of transcript per million.

"People's minds aren't made for problems that large."

Tyrion Lannister - Game of Thrones, "The Queen's justice"

3

Development of strategies and pipelines for analysing NGS data

3.1 Introduction

This Chapter addresses Aim 1 of this thesis; develop pipelines for handling large cohorts of next-generation sequencing data for gene discovery in ALS. The sheer magnitude of the data produced by next-generation sequencing (NGS), particularly whole-genome sequencing (WGS), poses a significant barrier to its effective use and interpretation. While there are a plethora of bioinformatics tools available to manipulate variant call files (VCFs), even these encounter difficulties when processing exceptionally large files, and the major challenge lies in determining how to effectively apply these tools to these large datasets. Therefore, this thesis involved the development of a range of strategies and pipelines to obtain the most robust and meaningful results from NGS data. These were developed to both manipulate VCFs to prepare them for genetic analysis, carry out the genetic analyses themselves and to also efficiently extract important data from analysed files to interpret their biological significance.

Whole-exome sequencing (WES) identifies approximately 80,000 variants in each

individual, while WGS identifies three to four million variants. Each of these variants has a range of associated variables, relating to genomic location, genotype and sequencing quality. When multiplied together, this equates to over 100 megabytes, and more than one gigabyte being required to represent an individual exome and genome, respectively. Standard computing systems and softwares are not well equipped to handle this data, and large volumes of memory are necessary for its storage. This poses a significant barrier to the effective utilisation of the genetic information contained within these files.

In order to effectively utilise the genetic information stored in NGS data files, coding strategies are required, as standard text editing and spreadsheet softwares cannot handle such large files. As part of this thesis, the UNIX and R environments, coupled with either bash scripts or R scripts, were implemented to analyse NGS data files. This Chapter presents a range of strategies and pipelines developed as part of this project to obtain the most robust and meaningful results from NGS data. The scripts included in this Chapter were developed for general genetic analysis tasks, which have been applied throughout the subsequent Chapters of this thesis, and are also routinely utilised by our research group. Various other scripts have been developed as part of this project to execute specific components of genetic analyses for known (Chapter 4), candidate (Chapter 5) or novel (Chapters 6 and 7) ALS gene identification. Such scripts are presented in the relevant Chapter, and are described as Custom Scripts.

This Chapter is divided into four sections:
1. The variant call format.
2. Bioinformatics tools and programs used in this project.
3. Development of scripting strategies to achieve vital manipulation of NGS data to facilitate efficient genetic analysis.
4. Development of complex pipelines to circumvent the difficulties encountered while handling large NGS datasets.

3.2 Variant call file format

A variant call file (VCF) is a tab-delimited file produced by standard bioinformatics processing of raw NGS reads (described in Chapter 2, Section 2.2.2.1) which describes each identified genomic variant (either a single nucleotide variant (SNV), or small insertion or deletion (indel)). The structure of a VCF is complex, and was designed

to be interpreted by both computers and people. An example of the VCF format is shown in Figure 3.1.

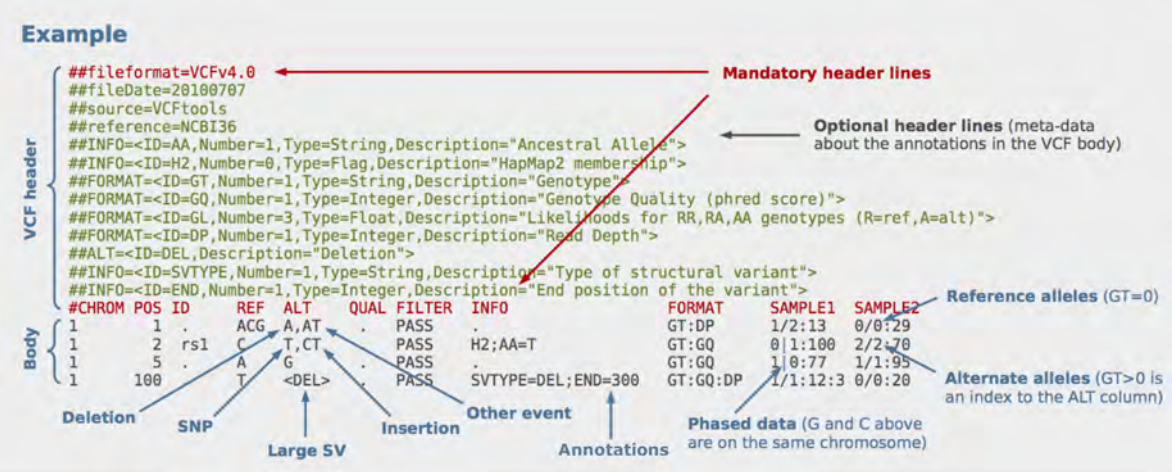


Figure 3.1: **Example of the VCF (variant call file) format.** The top lines consist of header information. The first header line invariably defines the VCF version, and the last header line is a classical header line containing descriptions of the contents of each column. The intervening header lines contain information about the processing and annotation steps applied to the variant data. Each line in the body of a VCF represents a genomic variant and details the corresponding meta information in separate columns including chromosome (CHROM), genomic DNA position (POS), identity (ID; if one exists), reference (REF) and alternate (ALT) nucleotide alleles. Quality information for each variant is found within the QUAL and FILTER columns. The INFO column can contain any number of biological annotation fields from various sources. In some instances, this INFO column can be separated out to individual columns for each annotation field. The FORMAT column describes what values are shown in each of the following sample data columns, of which there may be one or hundreds. Adapted from Danecek et al. (2011).

VCFs form the basis of a plethora of downstream genetic analyses. Typically, a WES VCF generated for a single person contains approximately 80,000 variants, while this figure sits in the vicinity of three to four million for WGS. These numbers equate to file sizes of approximately 125 and 1,500 megabytes for a single whole exome or genome respectively. Table 3.1 depicts the differences between VCFs generated from each of these NGS methods. Owing to this large file size, VCFs are not amenable for use with standard computing systems or programs, therefore specialised tools are necessary to analyse and manipulate these files.

Table 3.1: **Details of the size of VCFs produced from WES and WGS, in terms of memory and number of genetic variants.**

NGS method	Individual sample		Cohort*	
	VCF size	**No. of variants**	**VCF size**	**No. of variants**
WES	~125MB	~80,000	1.78GB	398,133
WGS	~11.5GB	~3,500,500	1.05TB	42,544,274

*WES cohort consists of FALS patients and any informative family members (n=137), and WGS of multiple cohorts (n=850) including SALS patients, FALS families, *SOD1* patients, FTD patients and ALS-discordant MZ twins. See Chapter 2, Figure 2.1 and Table 2.1 for details.

3.3 Computing and bioinformatics tools used for NGS data analysis and manipulation

The genetic information contained within VCFs formed the basis of the various genetic analyses presented throughout this thesis, including annotation, statistical analyses and variant filtering to identify genetic variants contributing to the cause of ALS. Numerous bioinformatics tools exist for the interrogation of the genetic data contained within VCFs, and those used as part of this thesis are described below.

3.3.1 High performance computing cluster

In order to handle the volume of data produced by WGS and the accompanying memory requirements for processing, access to a high performance-computing cluster (HPCC) was required. A HPCC is a scalable, data-intensive system which uses commodity server clusters hardware coupled to system software (Middleton, 2011). This platform provides a distributed file storage system, a job execution environment, parallel application processing and programming tools (Middleton, 2011). Various command line based codes developed in this thesis were executed using the CSIRO HPCC, Pearcey, to meet the computing demands of the analysis. The Pearcey cluster system runs Linux, and comprises 230 servers with 128GB memory each, and 16 servers with 512GB memory each.

3.3.2 Shell scripting

As the most widely adopted operating system, the UNIX programming language was used to develop various custom shell bash scripts to parse and manipulate text type files, including VCFs. A number of software tools as shown in Table 3.2 were also implemented as part of shell scripts using the UNIX environment.

3.3.3 R programming language

R is a programming language designed for statistical computing which provides a suite of capabilities for data manipulation, calculation and graphical display (R Core Team, 2018). It contains various tools for data analysis, while numerous extension packages have been developed to meet more specialised analytical needs (R Core Team, 2018). Bioconductor provides a suite of tools implemented in the R environment for analysing and interpreting genomic data (Huber et al., 2015). The R packages used in this project are summarised in Table 3.3.

TABLE 3.2: **Software programs utilised in this thesis using the UNIX environment.**

Software/Utility	Description	Scripts utilised	Reference
Basic UNIX	Basic commands of the UNIX coding language compatible with all delimited file types.	2.1, 3.1, 3.2, 3.4, 3.6, 3.7, 3.8, 3.9, 3.10, A.2.1, A.2.2, A.2.3.2, A.2.6, A.2.12, A.2.17	N/A
ANNOVAR	Software tool designed to utilise various databases for functional annotation of genetic variants. Annotations include gene-, region- and filter-based variant specific details. See Section 2.2.2.2 for further details.	2.1, A.2.1	Wang et al. (2010)
BCFTools	Program with various tools for the manipulation of VCFs and their binary counterpart file type, BCF.	3.6, 3.7, A.2.10, A.2.12	Li (2011)
Merlin	Pedigree analysis package.	A.2.17	Abecasis et al. (2002)
SnpSift	Program with many tools for filtering and manipulating annotated VCFs.	3.1, 3.4, 3.8, A.2.3.2	Cingolani et al. (2012)

Abbreviations: VCF, variant call file; and BCF binary variant call file.

Table 3.3: **R software packages utilised in this thesis.**

Package	Version/ release	Availability*	Use	Scripts used	Reference
Basic R	3.5.0	CRAN	Statistical computing language.	3.3, 3.5, 3.11, 3.12, 3.13, 3.14, A.2.3.1, A.2.3.2, A.2.4, A.2.5, A.2.8, A.2.7, A.2.9, A.2.11, A.2.13, A.2.14, A.2.15, A.2.16	R Core Team (2018)
data.table	1.11.4	CRAN	Data manipulations including subset, group and join.	A.2.4, A.2.8	Dowle and Srinivasan (2018)
dplyr	0.7.5	CRAN	Fast and consistent tool for working with data frame like objects.	A.2.4, A.2.18	Wickham et al. (2018)
gdata	2.18.0	CRAN	Data manipulations inlcuding operations and conversions.	A.2.7	Warnes et al. (2017)
ggplot2	3.0.0	CRAN	Creation of graphics.	A.2.18	Wickham (2016)
paramlink	1.1.2	CRAN	A suite of tools for analysing pedigrees with marker data.	A.2.15	Vigeland (2018)
readr	1.1.1	CRAN	Fast and friendly way to read rectangular data.	A.2.8, A.2.9	Wickham et al. (2017)
splitstackshape	1.4.4	CRAN	Splits concatenated data into separate cells.	A.2.3.2	Mahto (2017)
stringr	1.3.1	CRAN	Character manipulation and pattern matching.	3.13, A.2.8	Wickham (2018)
VariantAnnotation	1.26.0	Bioconductor	Annotation of genetic variants.	3.3, A.2.3.1, A.2.3.2	Obenchain et al. (2014)
WriteXLS	4.0.0	CRAN	Creation of Excel (xls and xlsx) files from dataframe objects.	A.2.4, A.2.5, A.2.8	Schwartz (2015)

Abbreviations: CRAN, The Comprehensive R Archive Network.

3.4 Development of basic scripts for NGS data processing, manipulation, and filtering

In order to conduct many downstream genetic analyses, an array of custom scripts were required to transform NGS data to the desired format for analysis, and subsequently perform data analysis and interpretation. This section outlines the major issues encountered when conducting genetic analyses on VCF data, and the codes developed as a solution for each such issue. Here, the most basic forms of these codes are provided, however, throughout the remainder of this thesis, various combinations have been utilised to produce files and results most appropriate for the given analysis. Most scripts were used for both WES and WGS data. The 850-sample WGS VCF (as described in Chapter 2, Figure 2.1 and Table 2.1) required analysis using the HPCC, while WES data and largely reduced formats of WGS data were analysed on standard computing systems.

3.4.1 Genotype quality filtering

Problem encountered

Numerous variants called by standard NGS processing were found to be sequencing artefacts following visualisation using the integrative genomics viewer (IGV; Robinson et al., 2011) and Sanger sequencing validation (see Figure 3.2 for an example). Genotype quality (GQ) filtering was applied (in hindsight) to subsequent analyses to reduce the frequency of such false positive variants being carried through analysis as candidate causal mutations.

Solution implemented

Variants with GQ scores less than 20 across all samples were removed from VCFs using the SNPSift *filter* tool, as shown in Custom Script 3.1. This approach was applied in Chapter 6, Section 6.2.2.1 for WES and WGS data.

Code 3.1: **GQfilter_VCF.sh** This script was developed using the SNPSift tool *filter*, and was used to filter a VCF based on genotype quality of the specified samples.

```
#!/bin/sh
#
#  GQfilter_VCF.sh

# remove any variant with a GQ value less than 20 for these three samples in the VCF
java -jar SnpSift.jar filter ' ( GEN[0].GQ > 20 ) & ( GEN[1].GQ > 20 ) & ( GEN[2].GQ > 20 ) ' EXAMPLE.vcf > EXAMPLE_GQfiltered.vcf
```

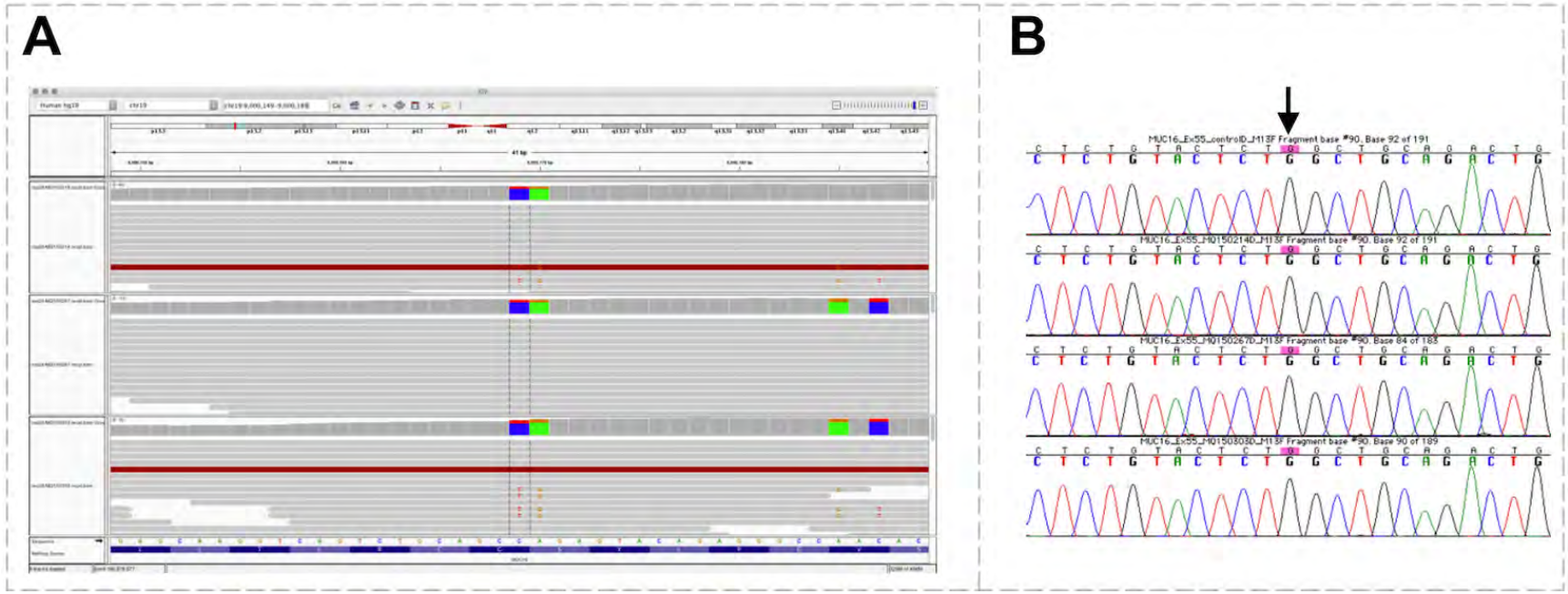


FIGURE 3.2: **Example of a false positive variant call identified using IGV and Sanger sequencing.** (A) IGV sequencing read visualisation. The variant was called as a heterozygous candidate mutation in WES data from three individuals of family FALSmq28 (two affected affected individuals and one obligate mutation carrier). Visualisation in IGV suggested that sufficient reads of each allele were present for a heterozygote call. (B) Sanger sequencing chromatograms. A non-related control individual and all three family members underwent Sanger sequencing for this variant, which showed that each had a homozygous wild-type genotype at this position, as indicated by the single peaks for each sample under the arrow.

3.4.2 Cutting and pasting VCF fields

Problem encountered

Circumstances often arose in which only certain VCF fields (for example AC, allele counts) were required for examination to conduct an effective analysis, while the efficiency of analysis would be greatly improved by only processing this smaller data subset. At times, the fields later needed to be combined back together, or with more detailed information.

Solution implemented

Three strategies were developed to meet this task. These codes were used interchangeably in Chapters 4, 5 and 6.

1) Custom bash scripts. As shown in Custom Script 3.2, the UNIX *cut* command was used to extract the columns of interest, and the UNIX *paste* command was used to combine these back together, side by side.

Code 3.2: **cut_paste_VCF.sh** This script utilised the UNIX *cut* command to extract data from a large VCF and create a new file containing a small subset. The UNIX *paste* command was then used to join these two subsets back together, horizontally.

```
#!/bin/sh
#
#  cut_paste_VCF.sh

# Take columns 1-9 for all lines after and including the line starting with a # symbol of a VCF, and write them to a new file
cut -f 1-9 EXAMPLE.vcf | grep -v -P '^#' > EXAMPLE_partA.vcf

# Take columns 836-838 and 842-844 for all lines after and including the line starting with a # symbol of a VCF, and write them to a new file
cut -f 836-838,842-844 EXAMPLE.vcf | grep -v -P '^#' > EXAMPLE_partB.vcf

# Take the two subsets, and join them back together horizontally
paste EXAMPLE_partA.vcf EXAMPLE_partB.vcf > EXAMPLE_partsAB.vcf
```

2) Custom R code implementing the R package VariantAnnotation from the Bioconductor suite. Firstly, the *scanVcfHeader* command was used to determine which fields were present in the VCF. These fields were then assigned as filtering parameters using the *ScanVcfParam* command. These parameters were then used in conjunction with the *readVcf* command to read only these fields into the R environment. This resulted in an S3 class object, a structure difficult for visual analysis. In order to obtain a simpler data format, the *rowRanges*, *mcols* and *info* commands were used to manipulate this object. This process is shown in Custom Script 3.3.

Code 3.3: **VCF_field.R** This script was used to extract one field of INFO data from a VCF. This example extracts the allele count filed.

```
# VCF_field.R

# load required R libraries
library(VariantAnnotation)
library(BiocInstaller)

# see what fields are present in this VCF
scanVcfHeader("EXAMPLE.vcf")

# define the paramaters on which we want to filter the VCF ie. alternate
    allele count
AC.param <- ScanVcfParam(info="AC_Adj")

# read this data into R studio
EXAMPLEvcf_AC <- readVcf("EXAMPLE.vcf", "hg19", param=AC.param) # s3 class
    object

# extract the the INFO column (AC and AN) data and genomic ranges
    information for each variant and combine
EXAMPLEvcf_AC_rowranges <- rowRanges(EXAMPLEvcf_AC)
mcols(EXAMPLEvcf_AC_rowranges) <- info(EXAMPLEvcf_AC)

# make this a data frame
EXAMPLEvcf_AC_rowranges_df <- as.data.frame(EXAMPLEvcf_AC_rowranges)
```

3) SnpSift. As shown in Custom Script 3.4, the *extractFields* tool from the SnpSift program was also used to extract data from specific INFO fields from a VCF, which were written to a new text file.

Code 3.4: **extractFields_VCF.sh** This script was used to extract the specified fields from a VCF and write these to a new file. This example extracts the fields for chromosome, position, reference and alternate alleles, and allele counts.

```
#!/bin/sh
#
#  extractFields_VCF.sh

# extract the chromosome, position, reference allele, alternate allele, alternate allele count and total allele counts from a VCF
java -jar SnpSift.jar extractFields EXAMPLE.vcf CHROM POS ID REF ALT AC NS > EXAMPLE_FieldsofInterest.txt
```

For Strategies 2 and 3, a unique variant identifying column in the format of "chr:position" was added to the resultant file using the R *paste* command. This column was then used as part of the R *merge* command in order to combine VCF fields together following separation, or to combine files from different sources by matching the "chr:position" columns. Custom Script 3.5 summarises this process.

Code 3.5: **merge.R** This script was used to add a variant identifying column to a VCF R dataframe in the form of "chr:position". It was also used to merge two R data frames based on matching the values contained in this identifying column.

```
# merge.R

# make a new column (chr:position) on the example dataframe, containing the contents of the CHROM and POS columns, separated by a ":"
example$chr.position <- paste(example$CHROM, example$POS, sep = ":")

# combine two dataframes by matching their values in the columns named "chr:position"
examples_combined <- merge( x=example1, y=example2, by.x="chr.position", by.y="chr.position", all.x = TRUE )
```

3.4.3 VCF comparisons

Problem encountered

When comparing analysis strategies for the same cohort, including ALS-discordant twins (Chapter 7), comparison of two or more VCFs was often required to determine the number and/or identity of common or unique variants.

Solution implemented

As shown in Custom Script 3.6, following VCF compression and indexing, the BCFTools *isec* command was utilised to create new VCFs of the overlapping (intersect) or unique (complimentary) variants between two or more VCFs.

Code 3.6: **isec_VCF.sh** This script was used to find the intersecting and complementary variants when comparing two or more VCFs.

```
#!/bin/sh
#
#  isec_VCF.sh

# compress the VCF
bgzip -c EXAMPLE.vcf > EXAMPLE.vcf.gz

# index the VCF
tabix -p vcf EXAMPLE.vcf.gz

# determine the variants that intersect between two VCFs and those that
    compliment each other
bcftools isec EXAMPLE1.vcf.gz EXAMPLE2.vcf.gz
```

3.4.4 Genomic region subsetting

Problem encountered

In some instances, only particular genomic regions were required for downstream genetic analysis. Therefore, these regions needed to be extracted from a VCF to a smaller subset VCF to facilitate efficient analysis. This was necessary to perform shared variant analysis on only those regions showing genetic linkage with disease in FALSmq28 (Chapter 6, Section 6.2.2.1), and to determine whether discordant variants between co-twins fell in confidently callable regions (Chapter 7, Section 7.2.1).

Solution implemented

A BED format-like text file defining the genomic regions to be analysed was written in a text editor program (example in Appendix A.4, Figure A.1). Using the Custom Script 3.7, this text file was converted to a UNIX readable format using the *dos2unix* command. Following compression and indexing, the BCFTools *view* command was used to create a new VCF containing only variants found to fall within the specified genomic region(s) contained in the text file that was defined by the R option.

Code 3.7: **subset_regions_VCF.sh** This script was used to create a subset of a VCF including only those variants falling within (a) given genomic region(s).

```
#!/bin/sh
#
#  subset_regions_VCF.sh

# convert the txt file to unix readable format
dos2unix -c Mac LOD_positive_regions.txt

# compress the VCF
bgzip -c EXAMPLE.vcf > EXAMPLE.vcf.gz

# index the VCF
tabix -p vcf EXAMPLE.vcf.gz

# generate a VCF (EXAMPLE_regionsONLY.vcf) that includes only the genomic
   regions specified in the file regions.txt
bcftools view -o EXAMPLE_regionsONLY.vcf -R regions.txt -Ov -s
   patient1,patient2,patient3 EXAMPLE.vcf.gz
```

3.4.5 Chromosomal splitting

Problem encountered

When analysing the VCF for the MGRB control dataset (details in Chapter 2, Table 2.4), downstream bioinformatics analysis was slow and tedious given the size of this VCF. As such, streamlining of downstream analysis required analysis by chromosome.

Solution implemented

The MGRB VCF was divided by chromosome using the SnpSift tool *SplitChr* to produce individual VCFs for each of the 22 autosomes, and the X and Y sex chromosomes (shown in Custom Script 3.8). Downstream analysis was then run individually on each chromosome VCF.

Code 3.8: **SplitChr_VCF.sh** This script was used to take a large VCF and create a single VCF for each chromosome.

```
#!/bin/sh
#
#  SplitChr_VCF.sh

# split VCF by chromosome
java -jar SnpSift.jar SplitChr HugeVCF.vcf.gz
```

3.4.6 VCF header

Problem encountered

When using UNIX *awk* commands to filter variants/lines from VCFs, header information was lost. In order to retain meaning of the resultant data, column headers needed to be reinstated. In special circumstances, such as ANNOVAR annotation, column headers were added for some fields, but remained missing for sample names.

Solution implemented

The column headers from an original VCF were extracted using the UNIX *grep* command. Subsequently, this header information was added back to the top of the processed VCF using the UNIX *cat* command, and where necessary, an incomplete header line was removed using the UNIX *sed* command. Custom Script 3.9 was developed for this purpose.

Code 3.9: **header_VCF.sh** This script was used to first create a single line file containing only the column header values from a VCF. This line was then added to the top of a processed VCF, and where appropriate, an incomplete header line introduced through processing was removed.

```
#!/bin/sh
#
#  header_VCF.sh

# Get the column header line from the original VCF by taking the line
    starting with the # symbol
grep '^#' EXAMPLE.vcf > HEADER.vcf

# Add the header line to the beginning of the annotated VCF
cat HEADER.vcf myanno_EXAMPLE.hg19_multianno.vcf >
    myanno_EXAMPLE.hg19_multianno_HEADED.vcf

# Remove the incomplete header line from the annotated VCF
sed -e '2d' myanno_EXAMPLE.hg19_multianno_HEADED.vcf >
    myanno_EXAMPLE.hg19_multianno_FINAL.vcf
```

3.4.7 Removing irrelevant variants

Problem encountered

When analysing WGS from just family FALSmq28, variants not called, or that were homozygous wild-type across all three family members were not interesting for analysis. Thereby, removal of these variants prior to analysis was necessary to substantially reduce computing requirements and analysis time.

Solution implemented

In order to remove variants, a UNIX *awk* approach was utilised to interrogate the three sample columns of interest to retain only variants that had an alternate allele call in at least one sample. That is, any variants that were either wild-type or not called in all three samples were removed. The Custom Script 3.10 was developed for this purpose.

Code 3.10: **filter_notcalled_homWT_VCF.sh** This script was used to remove any variants that had no called genotype or a homozygous wild-type genotype in all three individuals present in the VCF.

```
#!/bin/sh
#
#  filter_notcalled_homWT_VCF.sh


# remove all variants with no genotype called in all 3 individuals (present
    in columns 10-12)
awk ' ! ($10 ~ /^\.\/\.:/ && $11 ~ /^\.\/\.:/ && $12 ~ /^\.\/\.:/) {print
    $0}' EXAMPLE.vcf > EXAMPLE_called.vcf

# remove all variants with a homozygous WT genotype in all 3 individuals
    (present in columns 10-12)
awk ' ! ($10 ~ /^0\/0:/ && $11 ~ /^0\/0:/ && $12 ~ /^0\/0:/) {print $0}'
    EXAMPLE_called.vcf > EXAMPLE_called_noWThom.vcf
```

3.4.8 Variant filtering

Problem encountered

The most interesting variants were considered to be exonic, non-synonymous variants, as these affect amino acid sequence and therefore the encoded protein. Therefore, it was necessary to subset these exonic, non-synonymous variants for further analysis. This approach could also be extended to any other annotation of interest as required.

Solution implemented

Two R codes were used to retain only those variants that satisfied the set criteria. The first (Custom Script 3.11) was developed for use with VCFs for which the INFO column had been tab-delimited, and uses a simple R *which* command to define the rows to be retained based on the value of a column, and a second line using the R *subset* command to extract those lines from the complete file, while retaining header information. This version was used for WES and WGS subsets in Chapters 4, 5, 6 and 7. The second (Custom Script 3.12) was developed for use with VCFs where all INFO data, and therefore annotation data, were confined to a single column. This version used an R *grep* command to search within the INFO column, select and subset those lines containing the given value in that column. This version was used for WGS data generated from the 850-sample VCF in Chapters 4 and 5. The examples below were used to search for variants with an exonic function and non-synonymous annotation.

Code 3.11: **filter_variants_anno.R** This script was used to remove variants that did not match the given filtering criteria, by parsing the appropriate annotation column. This is an example for removing all variants that did not have a non-synonymous annotation in the exonic function column.

```
1 # filter_variants_anno.R
2
3 # retain only those variants with a "nonsynonymous" value in the
     ExonicFunc.refGene column
4 # this version was used for a VCF with a tab delimited INFO column
5 filter <- which(example$ExonicFunc.refGene == "nonsynonymous")
6 example.filtered <- example[filter,]
```

Code 3.12: **filter_variants_info.R** This script was used to remove variants that did not match the given filtering criteria, by parsing the INFO column containing annotation information. This is an example for removing all variants that did not have a non-synonymous annotation in the exonic function column.

```
# filter_variants_info.R

# retain only those variants with a "ExonicFunc.refGene=nonsynonymous"
    string in the INFO column
# this version was used for a VCF formatted file with a single INFO column
examplevcf_nonsynonymous <-
    example[grep("ExonicFunc.refGene=nonsynonymous", example$INFO), ]
```

3.4.9 Extracting variant annotations

Problem encountered

When analysing the 850-sample VCF, or a subset thereof, all the INFO data fell within a single column, containing up to hundreds of different annotations. As such, there was a need to efficiently extract the annotation of interest from the column for one, or many variants under investigation.

Solution implemented

The R command *str_match* from the stringr package was used to achieve this, by searching for a string of interest within the INFO column, and printing the output. The example in Custom Script 3.13 outputs the amino acid change for all variants in the VCF under analysis.

Code 3.13: **string_match.R** This script was used to extract specified annotation information from the INFO column for all variants in a VCF.

```
# string_match.R

# load required R libraries
library(stringr)

# What is the amino acid change for each variant?
x <- str_match(example$INFO, "AAChange.refGene=(.*?);")
x[,2] # this will print the AA change to the R console
```

3.4.10 Identifying samples containing a variant

Problem encountered

After filtering, the identity of the individual sample with a given genotype for those variants that remained under analysis, needed to be determined.

Solution implemented

A function was written in R using the *which* and *grepl* commands to find all columns, and therefore samples, matching the given genotype (Custom Script 3.14).

Code 3.14: **which_samples.R** This R function was used to determine which samples had (or did not have) a given genotype by outputting the column names of those columns matching the criteria.

```
# which_samples.R

# which samples (columns) do not have a homozygous wild-type or uncalled
    genotype?
which(apply(example[1,], 2, function(x) any(!grepl("0/0|\\./\\.", x))))
```

3.5 Pipelines developed and implemented for custom NGS processing

In order to conduct more complex and specialised processing of NGS data, pipelines involving multiple bioinformatics processing steps were developed. Frequently, this involved combining several of the codes described above into a larger pipeline script. Many such pipelines are presented in the following Chapters to address the specific needs of the relevant genetic analysis. The following sections detail three pipelines which were broadly applicable to multiple genetic analyses throughout this thesis, and within the research program underway in our laboratory.

3.5.1 ANNOVAR annotation of 850-sample WGS VCF

Problem encountered

As detailed in Chapter 2, Section 2.2.2.2, variant annotation is vital for filtering and interpretation of genetic data. Bioinformatics tools such as ANNOVAR (`http://annovar.openbioinformatics.org/en/latest/`; Wang et al., 2010) have largely solved the need for efficient and high throughout annotation. However, given exceptionally large VCFs, these tools have limitations. Initial attempts to use standard ANNOVAR codes to annotate the 850-sample VCF, which contains a total of 42,544,274 distinct variants equating to more than 1TB of data, failed due to the size of the file.

Solution implemented

In order to annotate the exceptionally large 850-sample VCF, the Custom Script in Appendix A.2.1 was developed, according to the process described in Figure 3.3. This involved splitting the original 850-sample VCF, so that the first sample and associated meta information could be annotated, and pasting this annotated single sample VCF back together with sample information for all other affected individuals. Application of this pipeline resulted in an 850-sample VCF successfully annotated with the databases listed in Chapter 2, Table 2.2.

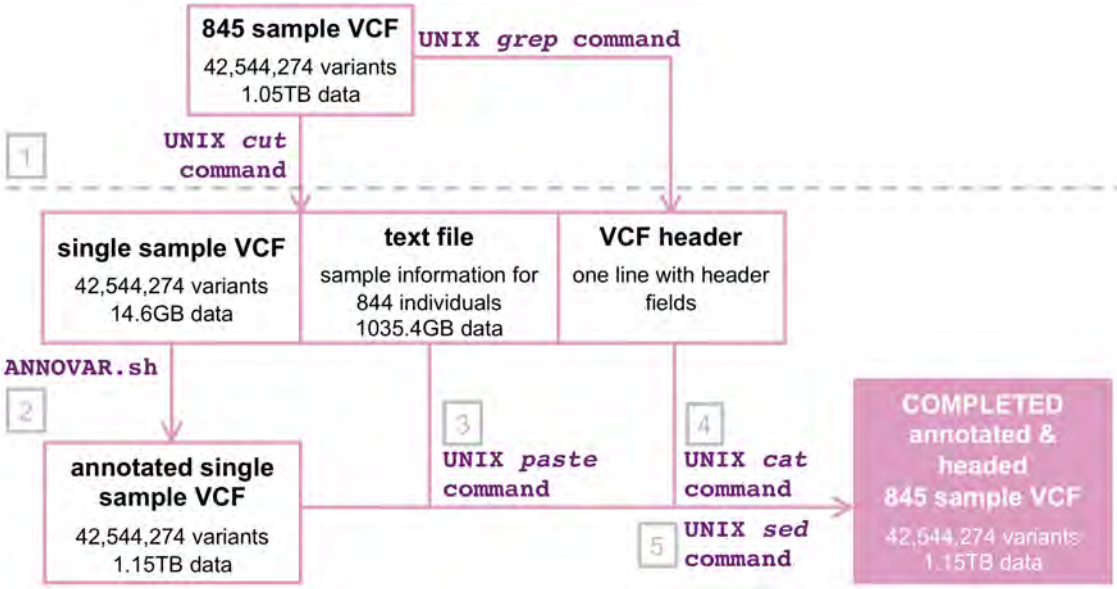


Figure 3.3: **Bioinformatic pipeline developed to annotate the 850-sample WGS VCF.** To overcome VCF size, the HPCC was used to execute the following steps to perform ANNOVAR annotation on the 850-sample WGS VCF. 1) Firstly, the meta information and the first sample were subset to a smaller, single sample VCF (UNIX *cut* command), while the sample information for all other individuals was subset to a separate text file (UNIX *cut* command). A column header line was also extracted from the original VCF (UNIX *grep* command). 2) The single sample VCF was annotated with ANNOVAR, using standard ANNOVAR code (Script 2.1). 3) The resultant annotated VCF was then combined with the sample information for all other individuals using a UNIX *paste* command. 4) Column header information was then added back to the resultant annotated file (UNIX *cat* command). 5) Finally, the incomplete header line added by ANNOVAR processing was removed (UNIX *sed* command).

3.5.2 Family subsetting from 850-sample WGS VCF

Problem encountered

ALS families present within the 850-sample WGS VCF required separate analysis pipelines, particularly in the case of family FALSmq28 (Chapter 6, Section 6.2.2.1), and needed to be separated from all other samples. Additionally, variants that were uninformative for all family members (i.e. variants not called or homozygous wild-type in all family members) required removal before efficient shared variant analysis was possible.

Solution implemented

Families were extracted from the complete 850-sample VCF using the Custom Script in Appendix A.2.2, which is summarised in Figure 3.4. This resulted in the creation of a three-sample VCF containing informative variant information for all three FALSmq28 family members utilised for ALS gene discovery in Chapter 6, Section 6.2.2.1.

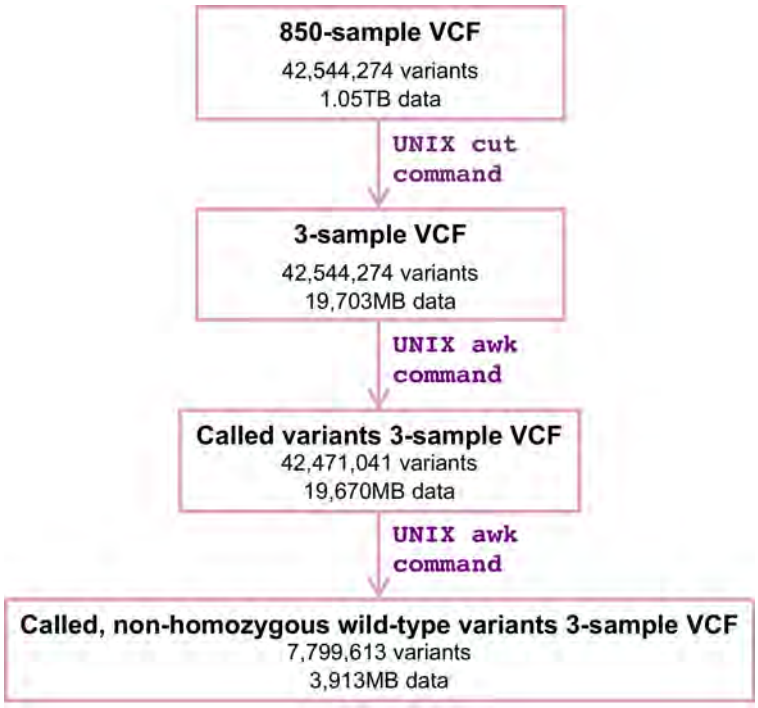


FIGURE 3.4: **Bioinformatic pipeline developed to subset families from the 850-sample WGS VCF.** To subset three family members from the 850-sample VCF, the UNIX *cut* command was used to write the meta information (columns 1-9) and the sample information for these three individuals (columns 836-838) to a new file. Variants that were not called in all three family members were then removed, followed by removal of homozygous wild-type variants in all three family members, using UNIX *awk* commands.

3.5.3 High-throughput analysis using publicly available control cohorts

Problem encountered

To accurately identify potential ALS causal mutations or disease associated SNPs, large cohorts of population- and age-matched controls must be screened. This was imperative to remove common variants as potential causes of ALS (Chapters 4, 5 and 6) and establish relative allele frequencies in the general population (Chapters 4 and 5). In recent years, NGS data from large cohorts of control individuals have been deposited into various publicly available databases (described in Section 2.4.3, Table 2.4). While invaluable resources, searching these databases for variant information is primarily facilitated through a web browser, meaning high-throughput screening is tedious and time consuming.

Solution implemented

In order to conduct high-throughput analysis in control datasets, allele count data were extracted from the four control databases described in Table 2.4 and combined with ALS patient VCFs to facilitate downstream comparisons. The VCF for each database was downloaded, and allele count data were then extracted using one of two Custom Scripting strategies, being either an R-based pipeline using the VariantAnnotation package from the Bioconductor suite, (Appendix A.2.3.1, Figure 3.5A), or a SNPSift pipeline (Appendix A.2.3.2, Figure 3.5B). In both cases, the control allele count data were subsequently combined with patient data using Custom R Scripts (Appendices A.2.3.1 and A.2.3.2, Figure 3.5C). The two different approaches were utilised as the first was developed in the early stages of candidature, before the SNPSift tool utilised in the second approach had become available. Application of these strategies resulted in the successful appendage of allele count data from the ExAC, gnomAD, DACC and MGRB control databases to the 137-sample FALS WES VCF, family WES VCFs for FALSmq28, FALS15, FALS45, FALSmq2 and FALSmq20 and gene and/or cohort subsets of the 850-sample WGS VCF. This enabled either variant filtering (Chapters 6, 5, 4) or association analysis (Chapters 4, 5).

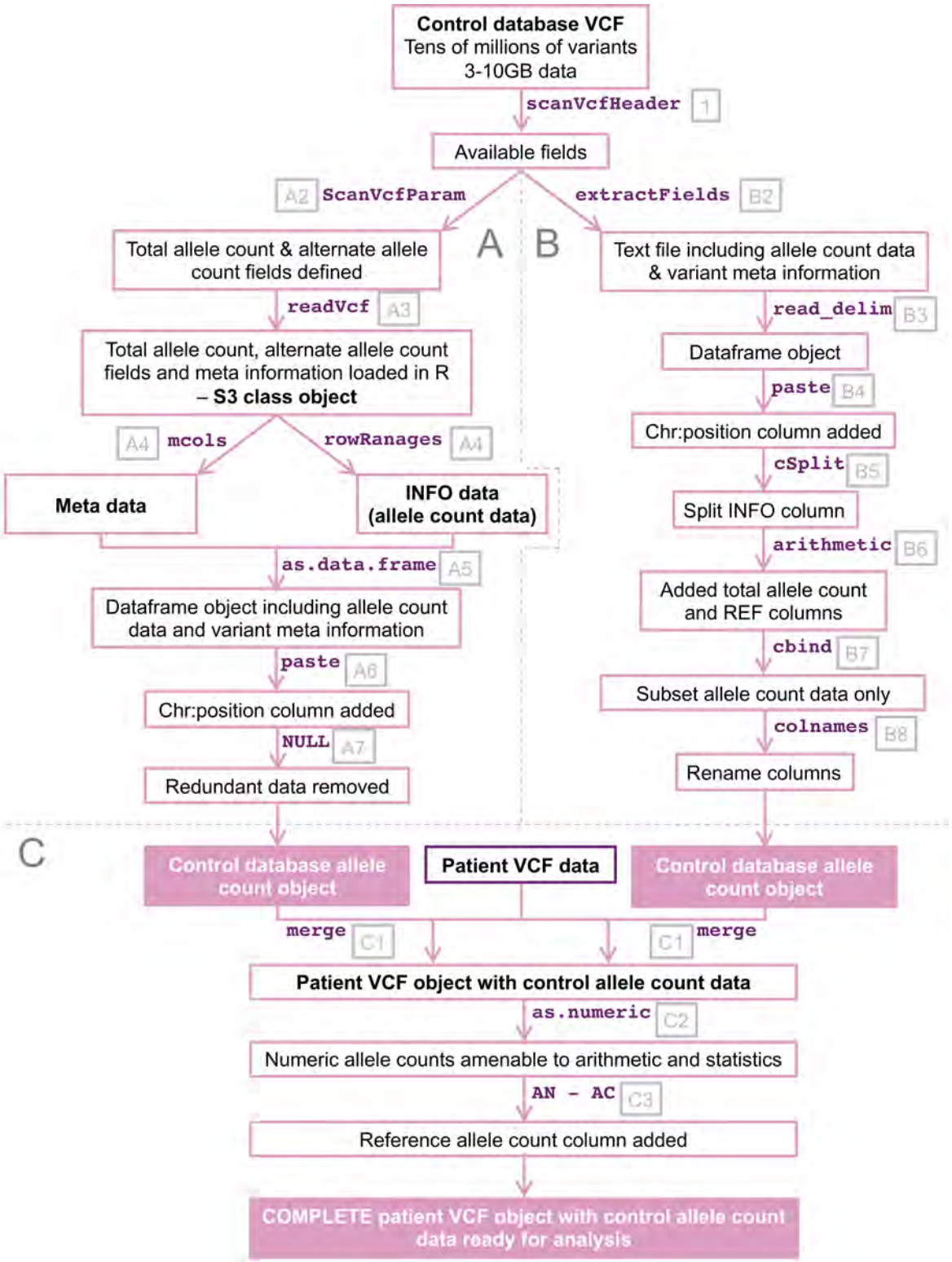


FIGURE 3.5: **Bioinformatic pipeline developed to extract and append control database allele count data to patient VCFs.** Caption provided on next page.

Figure 3.5: Two different approaches were employed to extract allele count data from control databases. 1) The first step of both approaches was to use the *scanVcfHeader* command to determine which VCF INFO fields were available, and what their associated identifiers were. A) R-based approach. This was applied for the ExAC, gnomAD exomes, gnomAD genomes and Diamentina VCFs. The fields of interest, being the alternate allele count (AC) and total allele count (AN) were defined as parameters using the *ScanVcfParam* command (A1). In conjunction with the *readVcf* command, these parameters were used to read in just these INFO fields of the control database VCF, and associated meta information (A2). This resulted in an S3 class object, a structure that is difficult for downstream visualisation. As such, this object was coerced to a more user friendly table like format (a data frame) by extracting the row (or INFO field ie. AC and AN) information, using the *rowRanges* command, and the meta information, using the *mcols* command (A3), and combining these together to make a user friendly data frame object (A4). For simplicity, and to facilitate downstream matching, a chromosome position column in the format of "chr:position" was created using the R *paste* command, to assign this unique identifying value for each variant (A5). Any redundant information was then removed by deleting unnecessary columns using the *NULL* command (A6), resulting in a data frame containing allele count data for each variant present in the given control database. B) SnpSift approach. This was applied to the MGRB VCF. The SnpSift *extractFields* command was used to write the desired meta information and allele count INFO fields to a new text file (B1), which was then imported into R as a data frame (B2), and also had the "chr:position" column added (B3). The dataframe included all INFO data in a single column, which was split into separate columns using the *cSplit* command from the R package splitstackshape (B4). Arithmetic functions were then applied to add a reference allele count column (B5). Relevant allele count data were combined to a new dataframe using the *cbind* command (B6), and columns were renamed using the *colnames* command (B7). C) Resultant data frames produced by either approach A or B were then used to append control allele counts to patient VCFs for comparison. After import to the R environment, the patient file also had the "chr:position" identifying column added (as described above). Based on matching the values in the "chr:position" column from both the patient and control data frames, the *merge* command was used to combine the two data frames (C1). Allele counts were then coerced to numeric values to facilitate use in arithmetic and statistical tests using the *as.numeric* command (C2). The reference allele count for each variant present in the control database was then added to a new column by subtracting the AC from the AN (C3). This produced a patient VCF data frame with appended control allele count data for downstream comparison.

3.6 Discussion

This Chapter has presented various custom scripting strategies and pipelines required for genetic analysis of NGS data. These scripts were developed to meet the specific needs of routine genetic analysis tasks performed in our laboratory for gene discovery in ALS patient cohorts. Prior to the development of these scripts, NGS data analysis was a tedious and time consuming process. Now, through implementing these scripting strategies, NGS data analysis is a streamlined and efficient process in our research team.

Only in hindsight did the need for quality-based filtering of NGS data become apparent. As seen in Chapter 6, Section 6.3.2 and Chapter 7, Manuscript III, numerous apparent candidate mutations were identified from NGS based genetic analysis. However, many of these were found to be false positive variant calls following Sanger sequencing validation. Not surprisingly, many of these false positive variants were identified within repetitive or duplicated genomic regions, which are notoriously difficult to accurately align to the reference genome (to be discussed in Chapter 8, Section 8.3.3). Further, many were also indel variants, that have a reputation for being particularly difficult to call, again due to incorrect read alignments (also to be discussed in Chapter 8, section 8.3.3). These indels were also often found within the aforementioned troublesome genomic regions, compounding the difficulties in calling these variants. Fortunately, the rate of identification of such variants could be substantially reduced by implementing genotype quality filtering. However, use of this filtering method introduces the potential to remove true variants from analysis. It was however a necessary step in order to produce a manageable number of candidate mutations from WGS data, and the likelihood of removing real variants was deemed to be exceptionally low.

Various strategies for minimising the volume of data under analysis have been presented in this Chapter. The use of techniques to reduce the number of samples and variants under analysis allowed for far more efficient downstream analysis, substantially reducing computing power and time requirements. These efforts also removed a large amount of irrelevant data from specific analysis, and therefore minimised the number of incidental findings that would have been difficult to interpret, causing inconclusive identification of variants of uncertain significance.

Extensive custom script writing was required to address specialised genetic analysis requirements throughout this project. Writing such Custom Scripts demanded a

significant time investment. First, learning various scripting languages is a lengthy process, with each having its own intricacies. Even after gaining an understanding of these languages, using them to write custom scripts to achieve a particular, complex purpose can often be difficult, especially when factoring in efficiency for use with large files. The process is largely trial-and-error based and relies on using smaller example datasets to initially develop a script, before use with larger files. Lines of code are particularly sensitive, and must be written completely accurately, or they may execute an entirely different function. Piping of one command to another can also have drastically different effects based on the order of analysis. Script development can therefore take weeks to perfect for a single purpose script, and also requires scrupulous error checking and trialling.

Interestingly, numerous strategies or pipelines utilising different tools can be employed to achieve the same processing goal. As a rapidly expanding field of study, and with constant advances in computing capabilities, it is not surprising that the release of new bioinformatics tools is a common occurrence. For instance, all three different approaches used for extracting fields from a VCF in Section 3.4.2 successfully achieved this task, however the intuitive nature of the different methods progressed. From basic column indexing in UNIX, to more complex field processing in R and finally to simple field definitions with SNPSift. Such advances in bioinformatics tools continue to make these analyses more accessible to biologists with minimal coding expertise.

In conclusion, bioinformatics processing and analysis of NGS data is difficult, and certainly presents a road block in any NGS study to effective interpretation downstream. However, a range of strategies and pipelines were developed in this Chapter to facilitate the use of this data. While the development of these approaches was time consuming and tedious, they were imperative for efficient and robust genetic analysis. As such, these strategies and pipelines have successfully been adopted to carry out ALS gene discovery in this thesis, and more broadly throughout many more aspects of the genomic ALS research program in our laboratory.

"Perhaps I was born with curiosity"

Panic! At The Disco - The Piano Knows Something I Don't Know

4

Analysis of known ALS genes

4.1 Introduction

This Chapter addresses the first part of Aim 2 of this thesis; to investigate known ALS genes in familial and sporadic Australian ALS patients to identify known and novel ALS mutations, and/or associated genetic variants. The purpose of the work in this Chapter was to assess the prevalence of established and recently reported ALS gene mutations, or disease associated variants, and identify any novel mutations in these genes in Australian familial (FALS) and/or sporadic (SALS) ALS patients. To achieve this, Sanger sequencing, bioinformatics analysis of next-generation sequencing (NGS) data, and high-throughput TaqMan genotyping was used. These analyses are presented as a collection of peer-reviewed publications/manuscripts, including one first author publication, one equal-first author manuscript, and three co-author publications to which the candidate significantly contributed to by screening ALS genes.

4.2 Methods

4.2.1 Sanger sequencing

Sanger sequencing of an ALS gene was performed by PCR amplification of genomic DNA (gDNA) from affected individuals, and subsequent Sanger sequencing. Details of this process are described in Chapter 2, Section 2.4.1, and primer information can be found in Appendix A.3, Table A.1.

4.2.2 NGS and bioinformatics analysis

Following the generation of NGS data as described in Chapter 2, Section 2.2, custom bioinformatics scripts were applied to the resultant variant call files (VCFs) to identify genetic variants in ALS patient sequencing data. When required, patient cohorts were extracted from this data using Custom Scripts (either that in Appendix A.2.4 or A.2.7). Gene screening was then executed using either the Custom Script in Appendix A.2.4 or A.2.6, for WES or WGS data respectively. Subsequently, any novel non-synonymous candidate mutations were identified using the Custom Scripts 3.11 or 3.12.

4.2.3 Custom TaqMan genotyping for association analysis

In order to determine whether a known population-based SNP was associated with Australian SALS, custom TaqMan genotyping was used to ascertain the frequency of the SNP among affected individuals and unrelated controls. DNA samples from control individuals collected at the Macquarie University Neurodegenerative Disease Biobank (n=108; 216 alleles) and the Australian MND DNA bank (n=535; 1070 alleles) were available for manual genotyping. These samples were screened for identified variants using custom high-throughput genotyping TaqMan assays (Life Technologies), to facilitate rapid and cost effective genotyping. Specific assay details are provided in Chapter 2, Section 2.4.3.2. Fisher's exact testing, with a significance threshold of 0.05, was then used to compare the number of alternate and wild-type alleles between affected individuals and controls.

4.3 Publications/manuscripts

4.3.1 Paper I – Screening and analysis of known ALS genes

ALS is a genetically heterogeneous disease, with at least 25 genes, and numerous individual mutations within each, known to cause disease (as established in Chapter 1, Section 1.4). It has also been noted that the frequency with which each ALS gene mutation causes disease differs between populations (discussed in Chapter 1, Section 1.4). In addition to this genetic heterogeneity, significant phenotypic heterogeneity is also evident amongst ALS patients. The clinical presentation of disease varies drastically in terms of age and site of onset, as well as rate of disease progression (discussed in Chapter 1, Sections 1.3.1 and 1.6.1). Particular ALS genes have also been associated with certain clinical characteristics, such as the predominance of *FUS* mutations in juvenile ALS (described in Chapter 1, Section 1.4).

In light of this heterogeneity, we endeavoured to describe the genotype-phenotype landscape of ALS in Australian FALS for the first time. This included determining the prevalence of each known ALS gene mutation, and identifying correlations between mutation status and either age of disease onset or disease duration. Additionally, we sought to determine the prevalence of pathogenic hexanucleotide repeat expansions in *C9orf72* among Australian SALS patients.

Two-hundred and twelve families, totalling 267 FALS patients, underwent comprehensive gene screening. Generally, FALS patients were first screened for the major ALS genes *C9orf72* and *SOD1* using repeat primed PCR and Sanger sequencing, respectively. Those patients negative for both then underwent WES. However, those FALS cases collected prior to the discovery of the pathogenic expansion of *C9orf72* in 2011 were screened for this locus in retrospect.

WES data was screened for any variants in ALS genes using the Custom Script in Appendix A.2.4, and any non-synonymous mutations were then identified using the Custom Script 3.11. This analysis showed that 60.8% of Australian FALS cases were explained by mutations in known ALS genes. Pathogenic expansion of *C9orf72* was evident in 40.6% of FALS, while ALS families harbouring mutations in *SOD1* (13.7%), *FUS* (2.4%), *TARDBP* (1.9%), *UBQLN2* (0.9%), *OPTN* (0.5%), *TBK1* (0.5%) and *CCNF* (0.5%) were also identified. Among the nine distinct *SOD1* missense mutations present in our cohort, p.V149G, p.I114T and p.E101G were most common, whilst p.A5V, the most frequent *SOD1* mutation in

North American, European-based populations (Andersen, 2006), was distinctly absent.

In order to investigate the relationship between age of disease onset or disease duration with the different ALS mutations, Kaplan-Meier survival analysis was performed. Subsequently, each of the different mutations were directly compared in a pair-wise manner using Mantel-Cox log-rank testing. A number of significant correlations were observed. Among the most interesting were that *C9orf72* expansion carriers were significantly more likely to develop disease later in life; the various *SOD1* mutations showed significant variance in clinical presentation of disease; and an apparent tendency of the *FUS* p.R521C mutation to show a more severe disease course than other non-*SOD1* mutations.

Further, since our laboratory's initial report on the Australian prevalence of pathogenic expansions in *C9orf72* (Williams et al., 2012b), a further 142 apparently sporadic patients had been recruited for our genetic studies. Initially, three of these SALS patients were found to carry the expansion, however, two were subsequently reclassified as familial cases following detailed genealogical analysis. This resulted in just 1/140 (0.7%) *C9orf72* expansion positive newly recruited SALS patients. When combined with the findings of our previous study (Williams et al., 2012b), just 2.9% of Australian SALS were found to carry a pathogenic expansion in *C9orf72*. Importantly, no phenotypic differences were observed between FALS and SALS expansion carriers, suggesting all *C9orf72* positive SALS may in fact be misclassified FALS patients.

Author contributions

The candidate performed all bioinformatics analyses; all statistical analyses; and all laboratory based gene screening on samples collected in or after 2014, and also wrote the manuscript. KW and JF performed laboratory based gene screening on samples collected prior to 2014, and provided intellectual input. IT provided intellectual input for statistical analysis. GN and DR collected samples and clinical information. JO also collected and collated clinical information. IB supervised the project and provided intellectual input. All authors contributed to the editing of the manuscript.

McCann, E. P., Williams, K. L.., Fifita, J. A., Tarr, I. S., O'Connor, J., Rowe, D. B., Nicholson, G. A. & Blair, I. P. (2017). The genotype–phenotype landscape of familial amyotrophic lateral sclerosis in Australia. *Clinical Genetics*, 92(3), p. 259-266.

DOI: 10.1111/cge.12973

Supplementary Table S1. The 22 pathogenic mutations identified in 212 Australian ALS families.

Gene	Transcript accession number	Nucleotide change	Amino acid change	First description of mutation by our laboratory
C9ORF72	NM_018325	Polymorphic g.26724GGGGCC(3_23)	NA	21
SOD1	NM_000454			
		c.19T>G	p.C7G	This report
		c.374A>T	p.D125V	This report
		c.272A>C	p.D91A	This report
		c.302A>G	p.E101G	This report
		c.217G>A	p.G73S	This report
		c.281G>T	p.G94V	This report
		c.131A>G	p.H44R	This report
		c.341T>C	p.I114T	This report
		c.446T>G	p.V149G	This report
FUS	NM_004960			
		c.1562G>A	p.R521H	20
		c.1562G>A	p.R521C	20
		c.1561C>A	p.R521S	This report
TARDBP	NM_007375			
		c.881G>T	p.G294V	19
		c.1009A>G	p.M337V	12
		c.1127G>A	p.G376D	18
		c.1158_1159delAT; c.1158_1159insCACCAACC	p.S387delinsTNP	18
OPTN	NM_ 001008211			
		c.883G>T	p.V295F	25
UBQLN2	NM_013444			
		c.1460C>T	p.T487I	22
TBK1	NM_013254			
		c.1197delC	p.L399fs	24
CCNF	NM_001761			
		c.1861A>G	p.S621G	23

Supplementary Table S2. Age of disease onset. Results of statistical analyses comparing ages of disease onset for ALS gene groups and *SOD1* variants. Statistically significant comparisons are highlighted in grey.

ALS gene groups (relates to figure 2A) Log-rank (Mantel-Cox) test: p<0.0001****		
Post-hoc Log-rank (Mantel-Cox) pairwise comparisons; Bonferroni corrected significance threshold: 0.0033		
Gene 1	**Gene 2**	**Log-rank (Mantel-Cox) p-value**
C9ORF72	*SOD1*	0.0003**
C9ORF72	*FUS*	<0.0001***
C9ORF72	*TARDBP*	0.0505
C9ORF72	*UBQLN2*	0.0002**
C9ORF72	*CCNF*	0.2780
SOD1	*FUS*	0.2108
SOD1	*TARDBP*	0.9705
SOD1	*UBQLN2*	0.1438
SOD1	*CCNF*	0.8598
FUS	*TARDBP*	0.6105
FUS	*UBQLN2*	0.6922
FUS	*CCNF*	0.5206
TARDBP	*UBQLN2*	0.3993
TARDBP	*CCNF*	0.9665
UBQLN2	*CCNF*	0.2511
Sample numbers: *C9ORF72*, n=222; *SOD1*, n=158; *FUS*, n=40 *TARDBP*, n=12; *UBQLN2*, n=15; *CCNF*, n=8.		
SOD1* mutations** (relates to figure 2D) Log-rank (Mantel-Cox) test: p<0.0001*		
Post-hoc Log-rank (Mantel-Cox) pairwise comparisons; Bonferroni corrected significance threshold: 0.005		
Variant 1	**Variant 2**	**Log-rank (Mantel-Cox) p-value**
p.H44R	p.E101G	0.9447
p.H44R	p.I114T	0.0207
p.H44R	p.D125V	0.0731
p.H44R	p.V149G	0.5454
p.E101G	p.I114T	0.0003**
p.E101G	p. D125V	0.1259
p.E101G	p.V149G	0.2598
p.I114T	p.D125V	0.6668
p.I114T	p.V149G	<0.0001***
p.D125V	p.V149G	0.0093
Sample numbers: p.H44R, n=10; p.E101G, n=27; p.I114T, n=69; p.D125V, n=6; p.V149G, n=41.		
Other gene mutations (relates to figure 2G) Log-rank (Mantel-Cox) test: p=0.0609		
Post-hoc Log-rank (Mantel-Cox) pairwise comparisons; Bonferroni corrected significance threshold: 0.008		
Variant 1	**Variant 2**	**Log-rank (Mantel-Cox) p-value**
FUS p.R521C	*FUS* p.R521H	0.0105
FUS p.R521C	*TARDBP* p.M337V	0.1103
FUS p.R521C	*UBQLN2* p.T487I	0.2329
FUS p.R521H	*TARDBP* p.M337V	0.3982

FUS p.R521H	*UBQLN2* p.T487I	0.0965
TARDBP p.M337V	*UBQLN2* p.T487I	0.6054

Sample numbers: *FUS* p.R521C, n=14; *FUS* p.R521H, n=24; *TARDBP* p.M337V, n=8; *UBQLN2* p.T487I, n=15.

Supplementary Table S3. Survival and disease duration. Results of statistical analyses comparing survival and disease duration for ALS gene groups and *SOD1* variants. Statistically significant comparisons are highlighted in grey.

ALS gene groups (relates to figure 2B)
Log-rank (Mantel-Cox) test: p=0.1572

Sample numbers: *C9ORF72*, n=117; *SOD1*, n=72; *FUS*, n=26 *TARDBP*, n=7; *UBQLN2*, n=14; *CCNF*, n=3.

***SOD1* mutations** (relates to figure 2E)
Log-rank (Mantel-Cox) test: p<0.0001 ****

Post-hoc Log-rank (Mantel-Cox) pairwise comparisons;
Bonferroni corrected significance threshold: 0.005

Variant 1	Variant 2	Log-rank (Mantel-Cox) p-value
p.H44R	p.E101G	<0.0001***
p.H44R	p.I114T	0.0021*
p.H44R	p.D125V	0.0771
p.H44R	p.V149G	0.0847
p.E101G	p.I114T	0.0929
p.E101G	p. D125V	<0.0001***
p.E101G	p.V149G	<0.0001***
p.I114T	p.D125V	<0.0001***
p.I114T	p.V149G	0.0088
p.D125V	p.V149G	<0.0001***

Sample numbers: p.H44R, n=6; p.E101G, n=14; p.I114T, n=23; p.D125V, n=5; p.V149G, n=21.

Other gene mutations (relates to figure 2H)
Log-rank (Mantel-Cox) test: p=0.0044**

Post-hoc Log-rank (Mantel-Cox) pairwise comparisons;
Bonferroni corrected significance threshold: 0.005

Variant 1	Variant 2	Log-rank (Mantel-Cox) p-value
FUS p.R521C	*FUS* p.R521H	0.0051
FUS p.R521C	*TARDBP* p.M337V	0.008
FUS p.R521C	*UBQLN2* p.T487I	0.0053
FUS p.R521C	*CCNF* p.S621G	0.0731
FUS p.R521H	*TARDBP* p.M337V	0.0675
FUS p.R521H	*UBQLN2* p.T487I	0.1605
FUS p.R521H	*CCNF* p.S621G	0.9621
TARDBP p.M337V	*UBQLN2* p.T487I	0.4316
TARDBP p.M337V	*CCNF* p.S621G	0.2562
UBQLN2 p.T487I	*CCNF* p.S621G	0.5806

Sample numbers: *FUS* p.R521C, n=9; *FUS* p.R521H, n=15; *TARDBP* p.M337V, n=5; *UBQLN2* p.T487I, n=14; *CCNF* p.S621G, n=3..

Supplementary Table S4. Age of disease onset and survival/disease duration. Results of statistical analyses comparing ages of disease onset as well as survival and disease duration between *FUS* p.R521C and *SOD1* mutations. Statistically significant comparisons are highlighted in grey.

Cumulative penetrance (relates to figure 3A) Log-rank (Mantel-Cox) test: p<0.0001****		
Post-hoc Log-rank (Mantel-Cox) pairwise comparisons; Bonferroni corrected significance threshold: 0.003		
Variant 1	**Variant 2**	**Log-rank (Mantel-Cox) p-value**
FUS p.R521C	*SOD1 p.H44R*	0.3845
FUS p.R521C	*SOD1 p.E101G*	0.3732
FUS p.R521C	*SOD1 p.I114T*	< 0.0001**
FUS p.R521C	*SOD1 p.D125V*	0.1186
FUS p.R521C	*SOD1 p.V149G*	0.5501
Sample numbers: *FUS* p.R521C, n=14; p.H44R, n=10; p.E101G, n=27; p.I114T, n=69; p.D125V, n=6; p.V149G, n=41.		
Survival and disease duration (relates to figure 3B) Log-rank (Mantel-Cox) test: p<0.0001****		
Post-hoc Log-rank (Mantel-Cox) pairwise comparisons; Bonferroni corrected significance threshold: 0.003		
Variant 1	**Variant 2**	**Log-rank (Mantel-Cox) p-value**
FUS p.R521C	*SOD1 p.H44R*	0.132
FUS p.R521C	*SOD1 p.E101G*	< 0.0001**
FUS p.R521C	*SOD1 p.I114T*	0.0267
FUS p.R521C	*SOD1 p.D125V*	< 0.0001**
FUS p.R521C	*SOD1 p.V149G*	0.6072
Sample numbers: *FUS* p.R521C, n=9; *SOD1* p.H44R, n=6; *SOD1* p.E101G, n=14; *SOD1* p.I114T, n=23; *SOD1* p.D125V, n=5; *SOD1* p.V149G, n=21.		

4.3.2 Manuscript II – Screening and analysis of *CHCHD10*, a newly reported ALS/FTD gene

CHCHD10 was first implicated in ALS by Bannwarth et al. (2014). A *CHCHD10* c.C176T; p.S59L mutation was identified in a French family with a mitochondrial DNA instability disorder. Many of the affected members of this family also exhibited a range of accompanying phenotypes, including symptoms suggestive of ALS and FTD. This led to genetic screening of *CHCHD10* in 21 ALS/FTD kindreds, which revealed an identical mutation was harboured by a proband of Spanish descent, thus affirming the pathogenic role of *CHCHD10* in ALS/FTD.

Following this initial discovery, mixed reports have surfaced concerning the contribution of genetic variation in *CHCHD10* to the cause of ALS/FTD in different populations. While many studies have successfully identified *CHCHD10* mutations in ALS/FTD cohorts (Dols-Icardo et al., 2015; Jiao et al., 2016; Johnson et al., 2014a; Kurzwelly et al., 2015; Muller et al., 2014; Perrone et al., 2017; Zhou et al., 2017), numerous others have reported their absence (Abdelkarim et al., 2016; Li et al., 2016; Marroquin et al., 2016; Teyssou et al., 2016; Wong et al., 2015). Interestingly, when considering these reports, it seems that *CHCHD10* mutations may be more frequent in FTD, or ALS/FTD patients than pure ALS cases. Additionally, some studies reporting *CHCHD10* variants as disease causal mutations have lacked comparison to large ethnically matched control cohorts (Chaussenot et al., 2014; Ronchi et al., 2015). Such claims have potentially contributed to an overestimation of the prevalence of *CHCHD10* mutations in ALS/FTD. As such, some studies have suggested that rare variants in *CHCHD10* may confer an increased disease-risk, rather than acting as ALS/FTD causal mutations (Abdelkarim et al., 2016).

Taken together, this led to the investigation of the prevalence of genetic variation in *CHCHD10* among Australian FALS, SALS and FTD patients. Additionally, we sought to investigate the pathology observed with the CHCHD10 protein in patient neurological tissue, as unlike many other ALS genes, this has not yet been reported.

The Custom Scripts in Appendices A.2.4 and A.2.6 were applied to WES data from FALS cases without a known causal mutation (n=81; including 61 probands), and WGS data from SALS patients (n=635) and FTD patients (n=108), respectively. Following analysis for non-synonymous variants using the Custom Scripts 3.11 or 3.12, three *CHCHD10* variants previously reported as beling linked to ALS and/or

FTD were identified. This included p.P34S, p.P80L, and p.P96T, each of which were present in six and two SALS cases, and one FTD case, respectively. However, all three variants are also present in multiple control individuals from the gnomAD database. Further, no novel *CHCHD10* mutations were identified. Thus we concluded that *CHCHD10* mutations are not a common cause of disease in the Australian population.

We also investigated whether any SNPs were associated with disease using the Custom Scripts in Appendices A.2.4 and A.2.9. While no SNPs were found to be associated with FALS or FTD, the p.P80L and p.P34S variants previously linked to ALS/FTD showed trends towards over-representation in SALS patients, compared with gnomAD and DACC controls, respectively. However, both trends were lost upon Bonferroni correction for multiple-testing. Interestingly, the initial reports of each of these variants in ALS/FTD had limited comparisons to control cohorts (Chaussenot et al., 2014; Ronchi et al., 2015), while both have subsequently been reported in both ALS/FTD patients as well as ancestrally-matched controls (Dobson-Stone et al., 2015; Wong et al., 2015), suggesting that they may not be true causal mutations.

In addition to the genetic analysis, we also sought to determine the localisation and expression of the CHCHD10 protein in patient neuronal tissue. Immunohistochemistry analysis of spinal cord and motor cortex tissues from ALS patients (including those with pathogenic *C9orf72* expansions or *SOD1* mutations, and SALS patients without a causal mutation) showed that CHCHD10 is expressed in motor neurons. While less CHCHD10 positive motor neurons were observed in patient spinal cord tissue, this was most likely due to a loss of motor neurons. As was established in Chapter 1, Section 1.3.4, the protein products encoded by numerous ALS genes have been found to aggregate as part of the disease hallmark protein inclusions observed in the affected motor neurons of ALS patients, and many also co-localise with TDP-43. As such, dual immunoflouresence staining was also completed for the spinal cord and motor cortex tissues from the aforementioned patients, to determine whether the CHCHD10 and TDP-43 proteins co-localised. While co-localisation was absent from the the spinal cord, some co-localisation was apparent in the motor cortex. Additionally, occasional CHCHD10-positive inclusions were seen in SALS patients.

Author contributions

The candidate performed all bioinformatics and genetic analyses including association analyses; performed IHC and IF experiments; assisted with microscopy and co-wrote the manuscript. JF provided intellectual input for genetic and association

analyses, performed IHC and IF experiments; performed microscopy and co-wrote the manuscript. JG, AH, NG, SF, SC and KZ performed IHC, IF experiments and microscopy. KW provided intellectual input for genetics analysis. NT and DB performed initial processing of the sequencing data used for bioinformatics analysis. PM and CJ performed western blotting and mouse tissue experiments and analysis. AW and JA supervised and provided intellectual input for western blotting and mouse tissue experiments. JK and GH provided FTD patient samples and clinical information. GN and DR collected ALS patient samples and clinical information. SY performed microscopy and image analysis; designed IHC, IF, western blot and mouse experiments; provided intellectual input for all pathology aspects; and contributed to writing the manuscript. IB supervised the project and provided intellectual input for all aspects. All authors contributed to the editing of the manuscript.

Genetic and immunopathological analysis of CHCHD10 in Australian amyotrophic lateral sclerosis and frontotemporal dementia

Corresponding author: Ian Blair, Centre for MND Research, Department of Biomedical Science, Faculty of Medicine and Health Sciences, 2 Technology Place, Macquarie University, Sydney, NSW, 2109, Australia. Email: ian.blair@mq.edu.au, phone: +61 2 9850 2725

Emily P. McCann[1*], Jennifer A. Fifita[1*], Natalie Grima[1], Jasmin Galper[1], Alison Hogan[1], Sarah Freckleton[1], Katharine Y Zhang[1], Sandrine Chan Moi Fat[1], Prachi Mehta[1], Cyril J Jagaraj[1], Kelly L. Williams[1], Natalie Twine[1,2], Denis Bauer[2], John Kowk[3,4], Glenda Halliday[3,4,5], Adam K Walker[1,6], Julie Atkin[1,7], Dominic B. Rowe[1], Garth A. Nicholson[1,8,9,10], Shu Yang[1^], Ian P. Blair[1^]

[1] Centre for MND Research, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia
[2] Commonwealth Scientific and Industrial Research Organization, Health & Biosecurity Flagship, Sydney, Australia
[3] Brain and Mind Centre, Sydney Medical School, The University of Sydney, Sydney, Australia.
[4] School of Medical Sciences, University of New South Wales, Sydney, Australia.
[5] Neuroscience Research Australia, Sydney, Australia
[6]Queensland Brain Institute, The University of Queensland, Queensland, Australia
[7] Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Bundoora, Melbourne, VIC 3086, Australia.
[8] Northcott Neuroscience Laboratory, ANZAC Research Institute, Sydney, New South Wales, Australia
[9] Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia
[10] Molecular Medicine Laboratory, Concord Hospital, Concord, New South Wales, Australia

* Equal first authors
^Equal senior authors

Keywords: amyotrophic lateral sclerosis, coiled-coil-helix-coiled-coil-helix domain containing 10 protein, neuropathology, neurogenetics

Word count: 4200
Number of tables: 1
Number of figures: 4

ABSTRACT

Objectives

Mutations in the *CHCHD10* gene have now been reported in mitochondrial DNA instability disorder, amyotrophic lateral sclerosis and frontotemporal dementia. To further assess the role of *CHCHD10* in ALS and FTD, we examined CHCHD10 pathology in motor and frontal cortex, and spinal cord tissues from ALS and FTD patients and controls. We also sought to determine the prevalence of *CHCHD10* mutations in Australian ALS and FTD.

Methods

Immunohistochemistry and immunofluorescence were performed to examine CHCHD10 localisation in spinal cord and motor cortex in a cohort of control and ALS patients, and frontal cortex in ALS-FTLD and pure FTLD patients. Western blotting was used to measure CHCHD10 expression in ALS patient motor cortex tissue and in cortex tissue from an inducible ALS TDP-43 mouse model. Mutation and association analysis of *CHCHD10* was performed by interrogation of whole exome and genome data from Australian FTD, and familial and sporadic ALS patients, as well as publicly available control databases.

Results

CHCHD10 showed primarily neuronal expression in spinal cord, motor cortex and frontal cortex tissues in both control and ALS patients. No significant changes were observed in CHCHD10 expression between control and ALS patients, but a significant downregulation of CHCHD10 was observed in the ALS TDP-43 mouse model following severe motor symptom onset. Three *CHCHD10* variants previously linked to ALS/FTD were identified in Australian ALS and FTD cases and controls. No novel mutations, or variant associations were identified.

Conclusions

We identified for the first time that CHCHD10 is localised primarily in the neurons of three central nervous system regions, suggesting a neuron-specific role for CHCHD10. CHCHD10 protein level changes are not evident in the motor cortex regions of ALS patient, but significant CHCHD10 reduction may be associated with disease progression in the ALS TDP-43 mouse model, suggesting potential interactions between CHCHD10 and TDP-43. This study also determined that *CHCHD10* mutations are not a common cause of FTD or familial and sporadic ALS in Australia.

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are two late-onset neurodegenerative diseases that are clinically, genetically and pathologically linked, and are therefore considered to sit along a spectrum of neurodegenerative disease.[1] ALS is characterised by the rapid degeneration of both upper and lower motor neurons, resulting in progressive muscle weakness, wasting, spasticity and eventual paralysis. Death generally occurs within two to five years of symptom onset.[2] FTD is characterised by progressive neuronal atrophy in the frontal and temporal cortices, leading to personality and behavioural changes. Up to 15% of ALS patients are diagnosed with comorbid FTD (ALS/FTD)[1], and 50% will develop some form of cognitive impairment.[1 3] In addition to this clinical overlap, ALS and FTD patients have a shared genetic aetiology, and both are characterised by the presence of protein aggregates in affected neurons.[1 3]

Approximately 10% of ALS patients (familial ALS; FALS) and more than a third of FTD patients exhibit familial inheritance of disease, with the remaining cases appearing apparently sporadically.[4-6] Both ALS and FTD are genetically heterogeneous diseases, where many genes with disease-causing mutations have been identified. These include the ALS-linked gene *SOD1,* [7] ALS/FTD-linked genes such as *TARDBP,*[8] *UBQLN2,*[9 10] *CCNF,*[11] and a hexanucleotide expansion within *C9orf72,*[12 13] and pure FTD-linked genes *MAPT*[14] and *PGRN*[15 16]. However, only two thirds of FALS, less than 10% of sporadic ALS (SALS)[17] and approximately 50% of familial FTD[18] patients carry known mutations, thus a significant number of genetic contributors to ALS as well as ALS/FTD are yet to be identified. Both ALS and FTD patients possess ubiquitinated neuronal cytoplasmic inclusions, which are the distinguishing pathological feature of both conditions[19 20]. These inclusions are positive for the TAR DNA binding protein, TDP-43 (encoded by *TARDBP*) in approximately 90% of ALS and 45% of FTD cases[21 22]. Pathological TDP-43 also undergoes characteristic biochemical changes, including hyperphosphorylation and cleavage[21 23].

In 2014, Bannwarth *et al*. [24] identified a novel mutation in the gene encoding the coiled-coil-helix-coiled-coil-helix domain containing 10 protein (*CHCHD10*) in a French family with a mitochondrial DNA instability disorder. The affected family members exhibited a range of accompanying phenotypes, including symptoms suggestive of ALS and FTD. Subsequently, an ALS/FTD patient from a family of Spanish descent was also found to carry an identical mutation. Since this initial report, numerous nonsynonymous *CHCHD10*

variants have been reported in ALS/FTD cohorts.[25-33] Some reports suggest that *CHCHD10* mutations are more closely linked to FTD dominant phenotypes than pure ALS. *CHCHD10* mutations have been found to be absent from multiple pure ALS cohorts,[34 35] and have appeared at much lower frequencies in ALS patients when compared to ethnically matched FTD cases.[26 35-38]

The function of CHCHD10 is largely unknown, though it is thought to be involved in mitochondrial organisation by interaction with the mitochondrial contact site and cristae organizing system (MICOS) protein complex. CHCHD10 interacts with the MIC60 protein complex (a component of MICOS) that is involved in inner mitochondrial membrane morphology[39]. A recent study by Woo et al.[40] identified a role for CHCHD10 in mitochondrial and synaptic integrity by identification of their dysfunction in a *Caenorhabditis elegans CHCHD10* (*har-1*) knockdown model. Additionally, transfection of the *CHCHD10* ALS/FTD mutations p.Arg15Leu and p.Ser59Leu into mouse primary neurons also induced abnormal mitochondria morphology, and reduced pre- and post-synaptic integrity, suggesting a loss-of-function toxicity. Interestingly, the study also found that CHCHD10 formed complexes with TDP-43, which promoted nuclear localisation of the protein, whereby expression of the ALS/FTD mutant CHCHD10 protein lead to cytoplasmic mislocalisation and aggregation of TDP-43 that co-localised with mitochondria.[40] However, nucleocytoplasmic translocation of TDP-43 was not observed in ALS patient-derived fibroblasts carrying a *CHCHD10* p.Gly66Val mutation.[41] Pathological studies of CHCHD10 in patient tissues are currently limited to the Bannwarth *et al*.[24] study, where patient muscle and skin fibroblasts from the family with mitochondrial DNA instability disorder, and a *CHCHD10* mutation, were analysed. Mitochondrial fragmentation, crystalloid inclusions and structural alterations were observed[24]. To-date, CHCHD10 pathology has not been investigated in ALS/FTD patient brain and spinal cord tissues.

This study set out to examine CHCHD10 pathology in ALS, ALS/FTD and FTD patient frontal cortex, and ALS patient spinal cord and motor cortex tissues. The study also aimed to determine the prevalence of *CHCHD10* mutations, or disease associated variants, in Australian ALS and FTD cohorts.

METHODS

Subjects and tissues

Eighty-one FALS patients (including 61 probands), and 628 SALS patients were ascertained from the Macquarie University Neurodegenerative disease Biobank, Molecular Medicine Laboratory at Concord Hospital, the Australian MND DNA bank, Royal Prince Alfred Hospital and the Brain and Mind Centre, University of Sydney. An additional 108 FTD patients were also recruited from the Brain and Mind Centre. All participants were recruited under informed written consent as approved by the human research ethics committees of the Sydney South West Area Health Service and Macquarie University, or Sydney University. Most participants were of European descent and ALS patients were all clinically diagnosed with ALS based on El Escorial criteria,[42] or FTD. Genomic DNA was extracted from peripheral blood using standard protocols.

Post-mortem paraffin-embedded human cervical spinal cord, motor cortex and frontal cortex sections (5μm), and fresh-frozen motor cortex tissue were obtained from the New South Wales Brain Bank Network and the Sydney Brain Bank. Cervical spinal cord tissues were available from SALS (n=10), *C9orf72* ALS (n=6), *SOD1* FALS (n=1), FALS (n=2) patients, and neurologically normal controls (n=5). Motor cortex tissues included SALS (n=11), *C9orf72* ALS (n=2), *SOD1* FALS (n=2) patients and neurologically normal controls (n=4). Frontal cortex tissues were available from sporadic frontotemporal lobar degeneration (FTLD)/ALS (n=3), *C9orf72* FTLD/ALS (n=3), *C9orf72* FTLD (n=3), sporadic FTLD (n=3) and neurologically normal controls (n=6). All FTLD and FTLD/ALS cases are characterised by TDP-43 inclusions.

Mouse cortex tissues were obtained from an established TDP-43 model of ALS (hTDP-43ΔNLS).[43 44] This model is characterised by ALS pathological features including accumulation of insoluble and phosphorylated cytoplasmic TDP-43 in the brain and spinal cord.

Patient tissue immunohistochemistry and immunofluorescence

Spinal cord, motor cortex and frontal cortex tissue sections underwent immunohistochemical analysis of CHCHD10. Spinal cord and motor cortex tissues also underwent dual immunofluorescence analysis of CHCHD10 and pathological phosphorylated TDP-43 (pTDP43). Tissue sections were pre-heated at 70°C for 30 min, were deparaffinized with xylene, and rehydrated with a descending series of ethanol washes. Antigens were retrieved by boiling sections in high pH buffer (pH 9.0, Dako, CA, USA) for 20 min. For immunohistochemical staining, endogenous peroxidase activity was

blocked using 3% hydrogen peroxide in methanol. Non-specific background was blocked using 5% normal goat serum (Vector Laboratories, CA, USA) with 0.1% Tween 20 in PBS for 1 h. Sections were incubated at 4°C overnight with primary antibodies: rabbit polyclonal anti-CHCHD10 (1:400, Sigma-Aldrich, MO, USA) alone for immunohistochemistry or in combination with mouse monoclonal anti-TDP-43 phosphorylated Ser409/410 (1:5000; Cosmo Bio, Japan) for immunofluorescence.

Sections were incubated at room temperature for 1 h with secondary antibodies: biotinylated goat anti-rabbit IgG (Vector Laboratories) for immunohistochemical staining and secondary alexaFluor-488 or 555 conjugated to anti-rabbit or anti-mouse antibodies (ThermoFisher Scientific, MA, USA) for immunofluorescent staining. For immunohistochemical staining, the avidin-biotin complex detection system (Vector Laboratories) with 3,3'-diaminobenzide as chromogen (Dako) was used to detect the immunoreactive signal. Immunohistochemistry sections were counterstained with hematoxylin and dehydrated with increasing series of ethanol washes followed by xylene. Sections were coverslipped using Di-N-Butyle Phthalate in xylene (DPX, Dako) or ProLong Gold antifade reagent with DAPI (ThermoFisher Scientific) for immunohistochemistry or immunofluorescence, respectively.

Visualisation and analysis of tissue sections

Immunohistochemistry sections were visualized using the ZEISS Axio Imager 2 microscope. Complete immunohistochemistry section images were captured using the Virtual Microscope ScanScope Unit and ScanScope Consol program before being visualised using the Image Scope program (Leica Biosystems, Germany). Immunofluorescence sections were imaged with a ZEISS LSM 880 inverted confocal laser-scanning microscope.

Generation of protein lysates from motor cortex tissue

Frozen motor cortex tissue was homogenized in 5X volume (µL/mg) of RIPA buffer (50mM Tris, 150mM NaCl, 1% Triton-X-100, 5mM EDTA, 0.5% sodium deoxycholate, 0.1% SDS, pH 8.0) containing phosphatase and protease inhibitors (Roche, Switzerland) using a motor-driven pestle. Homogenates were centrifuged at 124,500 x g for 40 min at 4°C. The supernatant was collected (RIPA-soluble fraction), and the pellet resuspended in 2X volume (µL/mg) of urea buffer (7M urea, 2 M thiourea, 4% CHAPS, 30 mM Tris, pH 8.5) containing phosphatase and protease inhibitors. The resuspensions were sonicated,

centrifuged at 124,500 x g for 40 min at 22°C and the supernatant collected. Protein concentration was determined using the Pierce BCA Protein Assay Kit (ThermoFisher Scientific).

Collection of mouse cortex tissues

hTDP-43ΔNLS or non-transgenic mice were deeply anesthetised using ketamine/xylazine and intracardially perfused with 15 ml phosphate-buffered saline followed by 30 min 10% formalin. Cortex tissues were dissected from four hTDP-43ΔNLS mice and four non-transgenic littermates at 2, 4 or 6 weeks off Dox.

Western blot analysis

Motor cortex protein lysates were prepared in dH_2O with Laemmli sample buffer (Bio-Rad) and NuPAGE sample reducing agent (ThermoFisher Scientific) and denatured at 70°C for 10 min. Protein lysates were electrophoresed into a 4-15% pre-cast polyacrylamide gel (Bio-Rad, CA, USA) and transferred to a nitrocellulose membrane using a semi-dry transfer (Bio-Rad, Trans-Blot Turbo Transfer System). Membranes were blocked in Odyssey Blocking Buffer in TBS (OBB) (LI-COR Biosciences, NE, USA) for 1 h at room temperature followed by overnight incubation at 4°C with primary antibodies: CHCHD10 (as above, 1:250), Neuronal Nuclei Antigen (NeuN) mouse monoclonal 1:1000 (Merck, Germany), TDP-43 rabbit polyclonal 1:2000 (Proteintech) or GAPDH mouse monoclonal 1:5000 (Proteintech). Membranes were then incubated for 1 h at room temperature with IRDye 680LT donkey anti-rabbit IgG and 800CW donkey anti-mouse IgG, 1:20,000 (LI-COR Biosciences). Antibodies were diluted in OBB with 0.1% Tween 20. Membranes were visualised using the Odyssey CLx imaging system and bands analysed with the Image Studio Lite software (LI-COR Biosciences).

Next-generation sequencing

FALS patients negative for mutations in *SOD1* and pathogenic expansions in *C9orf72* underwent whole exome sequencing (WES). Briefly, WES was performed at Macrogen Inc (Seoul, Korea) on the Illumina HiSeq2000 platform using the TruSeq Exome Enrichment kit (Illumina, CA, USA) or SureSelectXT Human All Exon V5 + UTR kit (Agilent, CA, USA). Full details of the cohort are described in.[17] SALS patients negative for the pathogenic expansion in *C9orf72* and FTD patients negative for mutations in *MAPT* and the *C9orf72* expansion, and two FTD patients with *GRN* mutations, underwent whole genome

sequencing (WGS) performed on the Illumina HiSeq X Ten platform using the TruSeq PCR-free library preparation (v2.5) (Kinghorn Centre for Clinical Genomics, Sydney, Australia). WES and WGS raw data were processed using the Genome Analysis ToolKit, (GATK, Broad Institute, MA, USA) and the corresponding best practices.[45-47] ANNOVAR[48] was used for annotation of variant call files (VCFs).

Genetic analysis

Using UNIX and the R statistical environment, custom bioinformatics analyses were applied to annotated VCFs to identify all genetic variants present in *CHCHD10* (NM_213720). The presence of *CHCHD10* variants previously reported as being ALS and/or FTD-linked was determined (c.44 G>T; p.Arg15Leu, c.100C>T; p.Pro34Ser, c.176C>T; p.Ser59Leu, c.197G>T; p.Gly66Val and c.239C>T; p.Pro80Leu, c.34C>T; p.Pro12Ser, c.244C>T; p.Gln82X, c.286C>A; p.Pro96Thr, c.67C>A; p.Pro23Thr, c.104C>A; p.Ala35Asp, c.64C>T, p.His22Tyr, c.68C>T, p.Pro23Leu, c.95C>A, p.Ala32Asp and c.170T>A, p.Val57Glu.) Variant alternate allele counts were compared between patients and unrelated control individuals using fisher's exact testing. Intergenic, upstream and downstream variants were not analysed. The p-value significance threshold was corrected for multiple-testing using Bonferroni corrections based on the number of variants identified. Patient allele frequencies were compared to three control datasets, the Non-Finnish European (NFE) WGS subset from the Genome Aggregation database (gnomAD),[49] and ethically matched control cohorts from the Medical Genome Reference Bank (MGRB, n=1144) and the Diamantina Australian Control Collection (DACC, University of Queensland, n=967). The MGRB and DACC cohorts consist of neurologically healthy individuals of predominantly Western European descent. Fisher's exact testing was not completed on variants absent from the MGRB and DACC data if flanking variants had low sequence coverage. An average of 15350 alleles was used to calculate p-values for variants absent in the NFE gnomAD control dataset.

RESULTS

Subcellular location of CHCHD10 in spinal cord, motor cortex and frontal cortex

We first performed IHC on ALS patient and control spinal cord and motor cortex tissues, as well as frontal cortex tissues from a small cohort of FTLD and/or ALS patients. In spinal cord (figure 1, A), CHCHD10 localised primarily in the grey matter region of both patients and controls. CHCHD10 expression was specifically observed in anterior horn motor neurons and neuropils and was generally absent from other cell types in both ALS patients

and controls. Less CHCHD10-positive motor neurons were visualised in ALS patient spinal cord sections compared to controls, likely due to motor neuron loss in ALS patients. A similar range of cytoplasmic expression was observed in controls and ALS patients with different genotypes (*C9orf72* repeat expansion, *SOD1* mutation and SALS with no known mutations, figure 1, B). In motor cortex and frontal cortex (figure 1, C-E), CHCHD10 showed cytoplasmic expression predominantly in medium and large pyramidal neurons located in cortical layers II, III and V in both patients and controls. A reduced number of CHCHD10 positive large pyramidal cells were seen in patients compared to the controls (figure 1, C, E). CHCHD10 location did not show a difference between control or ALS patients with different genotypes (figure 1 D, F). No difference was seen between ALS/FTLD and FTLD patients in terms of CHCHD10 localisation except that all three FTLD cases with a *C9orf72* repeat expansion showed low to no staining. To confirm the CHCHD10 location in these cases, additional sections were stained with an extended DAB incubation time (Supplementary figure 1). CHCHD10 expression was confirmed to be cytoplasmic in these cases after this staining, however one case still demonstrated low staining.

CHCHD10 did not form inclusions or colocalise with pTDP-43 inclusions in the majority of selected spinal cord and motor cortex tissues, and in all selected frontal cortex tissues (figure 2). CHCHD10 inclusions were observed in one or two neurons in spinal cord tissues from three SALS cases and motor cortex tissues from one SALS case (Supplementary figure 2).

CHCHD10 expression level in motor cortex

We observed variable levels of IHC staining both within the same case as well as between individual cases (Supplementary figure 3). Therefore, we sought to quantify CHCHD10 expression in control and patient tissues. Due to unavailability of fresh frozen spinal cord and frontal cortex tissues, Western blot analysis was conducted on motor cortex tissues only.

In motor cortex, CHCHD10 expression levels were examined by Western blot analysis of fresh frozen tissue lysates. Antibody specificity was confirmed by the presence of a single band product in line with previous findings[24]. Western blot analysis showed variable CHCHD10 expression between cases (figure 3, A). Since CHCHD10 localised primarily to the neurons in motor cortex regions, we used a neuronal marker, NeuN, to normalise

CHCHD10 expression levels. No significant changes were observed between control and ALS cases, except for one *SOD1* case which showed significantly higher expression than the control (figure 3 B, C).

We also examined CHCHD10 expression in an inducible ALS hTDP-43ΔNLS transgenic mouse.[44] We examined CHCHD10 protein level in transgenic or littermate non-transgenic control mice at 2, 4, or 6 weeks after removing suppressive reagent Dox, which corresponds to mild, medium and sever motor phenotypes. In motor cortex tissues from mice at six weeks off Dox, CHCHD10 expression is significantly reduced compared to control mice, but not in two or four week mice ($p<0.05$). NeuN expression did not show a significant difference between control and disease mice in any of these three time points (Supplementary figure 4).

***CHCHD10* variation in ALS and FTD**

Analysis of previously reported ALS and/or FTD-linked mutations

Whole exome and whole genome sequencing data was interrogated for the presence of *CHCHD10* variants in FALS and SALS/FTD respectively. Three disease-linked *CHCHD10* missense variants (c.100C>T; p.Pro34Ser, c.239C>T; p.Pro80Leu, and c.C286A; p.Pro96Thr) were present in six and two SALS cases, and one FTD case respectively (table 1). These three variants were also present in NFE gnomAD and MGRB controls (table 1). Interestingly, the ALS-linked variant, p.Pro80Leu, was absent from Australian controls and trended towards an overrepresentation in SALS compared to NFE gnomAD controls ($p=0.03$). However this was not significant after Bonferroni correction (as described below in the following section), nor was the trend replicated in DACC or MGRB controls (table 1). Additionally, a trend towards an overrepresentation of the ALS-linked p.Pro34Ser variant was seen in SALS when compared with DACC Australian controls ($p=0.0038$). Again, this was not significant after Bonferroni correction (as described below in the following section), nor was the trend replicated in gnomAD or MGRB control cohorts (table 1). One known rare nonsynonymous variant (c.T403C; p.Tyr135His, rs145649831) was also identified in SALS and FALS cases (table 1). No novel *CHCHD10* missense variants were identified.

Table 1. Nonsynonymous CHCHD10 variants identified in Australian ALS and FTD, and associated allele frequencies in cases and controls

Variant	ALS/FTD-linked variant	dbSNP ID	Cohort	AAF	GnomAD NFE		DACC		MGRB	
					AAF	p-value	AAF	p-value	AAF	p-value
c.100C>T, p.Pro34Ser	Yes	.	SALS	0.0048	0.004	0.651	0	0.0038	0.007	0.507
c.239TC>T, p.Pro80Leu	Yes	.	SALS	0.0016	0.0002	0.03	0	0.156	0	0.125
c.286C>A, p.Pro96Thr	Yes	rs111677724	FTD	0.0008	0.0008	1	0	0.101	0.0009	0.292
c.403T>C, p.Tyr135His	No	rs145649831	FALS	0.0068	0.0003	0.4219	0.0005	0.1395	0.0009	0.171
			SALS	0.0016	0.0003	0.065	0.0005	1	0.0009	1

Association analysis of population-based variants

Among FALS and SALS cases, a total of eight and 27 variants annotated as one of exonic, 3'UTR or intronic were identified in FALS and SALS cases respectively (Supplementary table 1). Therefore, the significance thresholds of $p<0.00625$ (FALS analysis) and $p<0.00185$ (SALS analysis) were applied after Bonferroni correction. Association analysis of population-based *CHCHD10* SNPs using Fisher's exact testing found no variants to be significantly associated with FALS or FTD (table 1, Supplementary table 1). One intronic SNP (rs62241575) was significantly associated with SALS compared to gnomAD NFE controls, however, analysis of Australian controls failed to replicate this association.

DISCUSSION

The current study identified CHCHD10 protein pathology in ALS and FTD patient tissues, and suggests the potential genetic contribution of *CHCHD10* in Australian ALS, ALS/FTD and FTD patients.

Mitochondrial dysfunction has long been recognised in ALS and FTD patients, however whether it is a cause or consequence of disease remains unclear. The recent identification of *CHCHD10* mutations in individuals within the ALS-FTD clinical spectrum has for the first time recognised genetic mutations in a mitochondrial protein as a potential cause of disease.[24] Since this discovery, studies attempting to elucidate the consequence of potentially pathogenic mutations have largely relied on skin fibroblasts from mutation carrying patients and overexpression of CHCHD10 mutants in *in vitro* and *in vivo* models. However, histopathological features of CHCHD10 in ALS and FTD cases without a CHCHD10 mutation have not been fully characterised.

In this study, we examined CHCHD10 localisation and expression levels in a set of neurologically normal controls, ALS, ALS/FTLD or FTLD patient post-mortem tissues. We found that CHCHD10 is primarily expressed in neurons of spinal cord, motor cortex and frontal cortex regions, and is generally absent from other cell types, such as glial cells, in both controls and patients. Neuronal mitochondria are highly dynamic organelles that are specialised in establishing and maintaining membrane excitability, neurotransmission and plasticity.[50] Our results suggest that CHCHD10 may have a neuron-specific role, possibly to support various neuronal activities as well as maintenance of mitochondrial network integrity. Furthermore, CHCHD10 mainly localised to the cytoplasm in both control and

patient neurons. Previous studies suggest CHCHD10 is a component of the mitochondria contact site and cristae organizing system complex.[39] Future studies are required to comprehensively illustrate the precise location of CHCHD10 in human motor neurons. We also observed CHCHD10 inclusion-like structures in a very small number of cases. It is unclear whether these inclusions are relevant for ALS pathogenesis. Future work on an extended cohort is required to determine the biological consequence of these structures.

While the initial study by Bannwarth et al.[24] did not report a significant reduction of CHCHD10 expression in *CHCHD10* mutant patient muscle tissue[24] several studies have since reported decreased CHCHD10 protein levels in *CHCHD10* mutation carrying patient derived fibroblasts and lymphoblasts compared to controls.[39 41 51 52] Such results favour the hypothesis that mutations in *CHCHD10* cause neurodegenerative disease via haploinsufficiency of *CHCHD10.* We did not observe significant CHCHD10 protein level changes between control and ALS patient motor cortex tissues, suggesting CHCHD10 protein level changes are not the primary cause of motor neuron death in these tissues. In contrast, CHCHD10 is downregulated in an inducible ALS TDP-43 transgenic mouse model (hTDP-43ΔNLS). A significant decrease in CHCHD10 protein levels was observed in mice at six weeks (presence of severe motor symptoms), but not at two (symptom onset and TDP-43 abnormalities) or four weeks (cortical atrophy and neuromuscular junction denervation) off the suppressive reagent Dox. This suggests that there may be an association between CHCHD10 reduction and disease progression.

It has previously been shown that TDP-43 can physically interact and form complexes with CHCHD10, while knockdown or expression of mutant CHCHD10 increases the accumulation of cytoplasmic TDP-43.[40] Our findings also clearly demonstrate an association between CHCHD10 abnormality and TDP-43-induced ALS pathogenesis in a model that is complimentary to the study by Woo et al., where TDP-43 is in a mutant form and CHCHD10 in its wild type. It also suggests that CHCHD10 changes may be downstream of the occurrence of TDP-43 abnormalities. Another interesting point is that all the ALS and FTLD cases used in study are characterised by TDP-43 pathologies, but yet CHCHD10 protein levels remain unchanged and CHCHD10 does not co-localise with pTDP-43 inclusions. Our interpretation is that perhaps CHCHD10 alteration only occurs in cells with severe TDP-43 pathologies. In our hTDP-43ΔNLS mouse model, TDP-43 showed significant biochemical changes such as an increased accumulation in RIPA-insoluble fractions compared to littermate non-transgenic controls.[44] In contrast, our ALS

cohort did not show any significant changes in RIPA-insoluble fractions between control and ALS patients (data not shown). Similarly, a previous study has utilised overexpression of either or both CHCHD10 and TDP-43 as a model to study the changes of these two proteins.[40] Therefore, in our cohort where both TDP-43 and CHCHD10 are at physiological levels and TDP-43 biochemical changes are mild, it may not be sufficient to induce CHCHD10 protein changes. It will be interesting to examine CHCHD10 expression specifically in neurons with severe TDP-43 pathologies or cases with TDP-43 mutations versus neurons with no or mild TDP-43 pathologies. Further histopathological studies are also warranted to identify whether CHCHD10 mislocalises or interacts with TDP-43 in this hTDP-43ΔNLS mouse model.

We found that three *CHCHD10* variants previously reported as ALS and/or FTD-linked (p.Pro34Ser, p.Pro80Leu, and p.Pro96Thr) were present in our large cohort of ALS and FTD patients. These variants were also found in control individuals, and their association with disease was not significant. One known intronic *CHCHD10* SNP (rs62241575) was found to be potentially over-represented in SALS compared with gnomAD NFE controls. However, this potential risk allele was not replicated using Australian control cohorts, suggesting its association is to Australian ethnicity rather than ALS. This finding highlights the critical importance of using ethnically matched control cohorts. Interestingly, we found that the p.Pro80Leu, variant reported to be pathogenic by Ronchi et al.[53], trended to overrepresentation in SALS patients compared to NFE gnomAD controls, however this too was lost upon comparisons with Australian control cohorts. Notably, in their initial report, Ronchi et al.[53] identified p.Pro80Leu in two SALS patients and found it to be absent from the 1000 Genomes and Exome Variant Server control databases and an additional 286 Italian controls. However, this equates to approximately 7,500 control individuals, whereas here we have utilised data from over 18,000 healthy individuals using the NFE gnomAD control cohort and two Australian control cohorts, providing far greater power to determine the novelty, or apparent disease association of genetic variants. The above findings, as well as results from Dobson-Stone *et al.* (2015),[54] reiterate that screening of large ethnically matched control cohorts is critical to accurately assess the pathogenicity of potential disease gene variants. Our results suggest that genetic variation in *CHCHD10* is not a common cause of, or risk factor for, ALS or FTD in Australia.

Altogether, we reported CHCHD10 location and expression in ALS and FTD post-mortem tissues for the first time. Our result suggests that CHCHD10 plays roles primarily in

neurons and CHCHD10 abnormality can be found in patients without CHCHD10 mutation. At this stage, it is not clear whether reduced CHCHD10 levels are the cause or the result of motor neuron degeneration or mitochondrial cristae dysfunction. Further efforts to investigate mitochondria cristae changes in ALS post-mortem tissues and ALS mouse models should provide more insights into its role in ALS pathogenesis. It will also be interesting to further elucidate the interaction between CHCHD10 and C9orf72 dipeptide repeats. The impact of the genetic findings reaffirms that while it appears genetic variation in *CHCHD10* does contribute to the aetiology of ALS, it may not always be as an autosomal dominant cause of disease, and may often be contributing to disease risk through interactions with a other genetic variations waiting to be uncovered.

ACKNOWLEDGEMENTS

The authors thank L. Adams, A. Crook, C. Cecere, and J. O'Connor for their assistance in sample collection and compiling family information, the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource (a full list of contributing groups can be found at http://gnomad.broadinstitute.org/about), Paul Leo, Emma Duncan and Matthew Brown for access to whole exome data from the Diamantina Australian Control Collection 1.0, the use of Australian WGS data generated by the MGRB Collaborative (http://sgc.garvan.org.au/mgrb/initiatives), and the New South Wales Brain Bank and Sydney Brain Bank for providing tissues.

DNA samples used in this research were obtained from the Macquarie University Neurodegenerative Disease Biobank, Macquarie University, New South Wales, Australia, Northcott Neuroscience Laboratory, Concord Hospital, MNDDNA bank, Sydney University, and Brain and Mind Centre, Sydney University.

COMPETING INTERESTS

None declared.

FUNDING

This work was funded by the Motor Neuron Disease Research Institute of Australia (Bill Gole Postdoctoral Research Fellowship to JAF, PhD scholarship top-up to EPM), MND Australia (Leadership Grant to IPB), and the National Health and Medical Research Council of Australia (1095215, 1092023). The Diamantina Control Cohort includes data obtained from projects funded by NHMRC Project Grants 1032571 and 511132.

FIGURE LEGENDS

Figure 1. IHC staining of CHCHD10 in (A, B) spinal cord, (C, D) motor cortex and (E, F) frontal cortex in human post-mortem tissues. A, C, E are ScanScope images of the whole section from one control and one patient from each location, and the zoomed-in view of the boxed area. B, C, D were neurons from control and patients with different genotypes taken with Zeiss Axio Imager using a 20x lens. A, C, E showed that in all three locations, CHCHD10 showed positive staining primarily in the grey matter, and strong immunoreactivity specifically to the neurons. Reduced numbers of CHCHD10 positive neurons were seen in all three locations (A, C, E, Boxed area). In all three locations, CHCHD10 showed primarily cytoplasmic localisation. No significant difference was seen in terms of subcellular location between control and patients, and between patients with different genotypes (B, D, E).

Figure 2. CHCHD10 did not colocalise with TDP-43 inclusions in spinal cord (A), motor cortex (B) or frontal cortex (C) tissues. TDP-43 inclusions were labelled with an antibody that is specific to phosphorylated TDP-43 (green) and CHCHD10 was labelled with anti-CHCHD10 antibody (red). Colocalisation between TDP-43 inclusions and CHCHD10 was not observed in most of the neurons, except for a small number of motor cortex neurons, where partial colocalisation was evident (B, insets). Images were photographed using a 63X lens. Scale bar: 20μm.

Figure 3. CHCHD10 showed variable expression in post-mortem motor cortex tissues. (A) Western blotting of control or ALS motor cortex tissues; (B) Semi-quantification of CHCHD10 expression normalised NeuN showed variable expression between individual cases. (C) Semi-quantification data of B was grouped as control and ALS. A decreasing trend was seen in ALS cases, although the difference is not significant between control and patient.

Figure 4. CHCHD10 expression in a mouse model of ALS (rNLS TDP-43 mice). Mouse brain (cortex) was collected at at pre-symptomatic (2 weeks) and symptomatic (4 & 6 weeks) TDP-43 mice and litter-matched controls (n=4/group) and was immunoblotted with CHCHD10 antibody. No significant changes in the expression of CHCHD10 was seen at 2 and 4 weeks post-disease onset. In contrst, a significant decrease in the expression of the CHCHD10 gene was observed at 6 week post-onset in diseased mice compared to

controls.

Supplementary Figure 1. IHC staining of CHCHD10 in Non-C9orf72 FTD frontal cortex tissues with long DAB exposure. Frontal cortex sections from three Non-C9orf72 FTD cases were incubated with DAB for 5 min. While Case 1 and 2 showed increased staining, Case 3 remained weak.

Supplementary Figure 2. CHCHD10 (green) formed dense dot inclusion-like structures in a few neurons from two SALS spinal cord cases and one SALS motor cortex cases.

Supplementary Figure 3. Examples of variable IHC staining levels within the same case and in between cases

Supplementary Figure 4. NeuN staining showed no significant difference between control and diseased mice at 2, 4 or 6 weeks off-Dox.

REFERENCES

1 Ringholz GM, Appel SH, Bradshaw M, et al. Prevalence and patterns of cognitive impairment in sporadic ALS. *Neurology* 2005;65(4):586-90.

2 de Carvalho M, Swash M. Amyotrophic lateral sclerosis: an update. *Curr Opin Neurol* 2011;24(5):497-503.

3 Montuschi A, Iazzolino B, Calvo A, et al. Cognitive correlates in amyotrophic lateral sclerosis: a population-based study in Italy. *Journal of neurology, neurosurgery, and psychiatry* 2015;86(2):168-73.

4 Renton AE, Chio A, Traynor BJ. State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* 2014;17(1):17-23.

5 Boylan K. Familial Amyotrophic Lateral Sclerosis. *Neurologic clinics* 2015;33(4):807-30.

6 Woollacott IO, Rohrer JD. The clinical spectrum of sporadic and familial forms of frontotemporal dementia. *Journal of neurochemistry* 2016;138 Suppl 1:6-31.

7 Rosen DR, Siddique T, Patterson D, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 1993;362(6415):59-62.

8 Sreedharan J, Blair IP, Tripathi VB, et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* 2008;319(5870):1668-72.

9 Williams KL, Warraich ST, Yang S, et al. UBQLN2/ubiquilin 2 mutation and pathology in familial amyotrophic lateral sclerosis. *Neurobiology of aging* 2012;33(10):2527 e3-10.
10 Deng H-X, Chen W, Hong S-T, et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature* 2011;477(7363):211-15.
11 Williams KL, Topp S, Yang S, et al. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nature communications* 2016;7:11253.
12 DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 2011;72(2):245-56.
13 Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 2011;72(2):257-68.
14 Hutton M, Lendon CL, Rizzu P, et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* 1998;393(6686):702-5.
15 Baker M, Mackenzie IR, Pickering-Brown SM, et al. Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature* 2006;442(7105):916-9.
16 Cruts M, Gijselinck I, van der Zee J, et al. Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature* 2006;442(7105):920-4.
17 McCann EP, Williams KL, Fifita JA, et al. The genotype-phenotype landscape of familial amyotrophic lateral sclerosis in Australia. *Clin Genet* 2017
18 Pottier C, Ravenscroft TA, Sanchez-Contreras M, et al. Genetics of FTLD: overview and what else we can expect from genetic studies. *Journal of neurochemistry* 2016;138 Suppl 1:32-53. doi: 10.1111/jnc.13622
19 Leigh PN, Dodson A, Swash M, et al. Cytoskeletal abnormalities in motor neuron disease. An immunocytochemical study. *Brain : a journal of neurology* 1989;112 (Pt 2):521-35.
20 Kovari E, Leuba G, Savioz A, et al. Familial frontotemporal dementia with ubiquitin inclusion bodies and without motor neuron disease. *Acta neuropathologica* 2000;100(4):421-6.

21 Neumann M, Sampathu DM, Kwong LK, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 2006;314(5796):130-3.
22 Ling SC, Polymenidou M, Cleveland DW. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* 2013;79(3):416-38.
23 Arai T, Hasegawa M, Akiyama H, et al. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and biophysical research communications* 2006;351(3):602-11.
24 Bannwarth S, Ait-El-Mkadem S, Chaussenot A, et al. A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain : a journal of neurology* 2014;137(Pt 8):2329-45.
25 Dols-Icardo O, Nebot I, Gorostidi A, et al. Analysis of the CHCHD10 gene in patients with frontotemporal dementia and amyotrophic lateral sclerosis from Spain. *Brain : a journal of neurology* 2015;138(Pt 12):e400.
26 Jiao B, Xiao T, Hou L, et al. High prevalence of CHCHD10 mutation in patients with frontotemporal dementia from China. *Brain : a journal of neurology* 2016;139(Pt 4):e21.
27 Johnson JO, Glynn SM, Gibbs JR, et al. Mutations in the CHCHD10 gene are a common cause of familial amyotrophic lateral sclerosis. *Brain : a journal of neurology* 2014;137(Pt 12):e311.
28 Kurzwelly D, Kruger S, Biskup S, et al. A distinct clinical phenotype in a German kindred with motor neuron disease carrying a CHCHD10 mutation. *Brain : a journal of neurology* 2015;138(Pt 9):e376.
29 Muller K, Andersen PM, Hubers A, et al. Two novel mutations in conserved codons indicate that CHCHD10 is a gene associated with motor neuron disease. *Brain : a journal of neurology* 2014;137(Pt 12):e309.
30 Perrone F, Nguyen HP, Van Mossevelde S, et al. Investigating the role of ALS genes CHCHD10 and TUBA4A in Belgian FTD-ALS spectrum patients. *Neurobiology of aging* 2017;51:177.e9-77.e16.
31 Shen S, He J, Tang L, et al. CHCHD10 mutations in patients with amyotrophic lateral sclerosis in Mainland China. *Neurobiology of aging* 2017;54:214.e7-14.e10.
32 Zhou Q, Chen Y, Wei Q, et al. Mutation Screening of the CHCHD10 Gene in Chinese Patients with Amyotrophic Lateral Sclerosis. *Mol Neurobiol* 2016

661 33 Zhang M, Xi Z, Zinman L, et al. Mutation analysis of CHCHD10 in different
662 neurodegenerative diseases. *Brain : a journal of neurology* 2015;138(Pt 9):e380.
663 34 Li XL, Shu S, Li XG, et al. CHCHD10 is not a frequent causative gene in Chinese
664 ALS patients. *Amyotroph Lateral Scler Frontotemporal Degener* 2016;17(5-6):458-
665 60.
666 35 Teyssou E, Chartier L, Albert M, et al. Genetic analysis of CHCHD10 in French
667 familial amyotrophic lateral sclerosis patients. *Neurobiology of aging*
668 2016;42:218.e1-3.
669 36 Chaussenot A, Le Ber I, Ait-El-Mkadem S, et al. Screening of CHCHD10 in a
670 French cohort confirms the involvement of this gene in frontotemporal dementia
671 with amyotrophic lateral sclerosis patients. *Neurobiology of aging*
672 2014;35(12):2884.e1-4.
673 37 Marroquin N, Stranz S, Muller K, et al. Screening for CHCHD10 mutations in a large
674 cohort of sporadic ALS patients: no evidence for pathogenicity of the p.P34S
675 variant. *Brain : a journal of neurology* 2016;139(Pt 2):e8.
676 38 Wong CH, Topp S, Gkazi AS, et al. The CHCHD10 P34S variant is not associated
677 with ALS in a UK cohort of familial and sporadic patients. *Neurobiology of aging*
678 2015;36(10):2908.e17-8. doi: 10.1016/j.neurobiolaging.2015.07.014
679 39 Genin EC, Plutino M, Bannwarth S, et al. CHCHD10 mutations promote loss of
680 mitochondrial cristae junctions with impaired mitochondrial genome maintenance
681 and inhibition of apoptosis. *EMBO molecular medicine* 2016;8(1):58-72.
682 40 Woo JA, Liu T, Trotter C, et al. Loss of function CHCHD10 mutations in cytoplasmic
683 TDP-43 accumulation and synaptic integrity. *Nature communications* 2017;8:15558.
684 41 Brockmann SJ, Freischmidt A, Oeckl P, et al. CHCHD10 mutations p.R15L and
685 p.G66V cause motoneuron disease by haploinsufficiency. *Human molecular*
686 *genetics* 2018;27(4):706-15.
687 42 Brooks BR, Miller RG, Swash M, et al. El Escorial revisited: revised criteria for the
688 diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor*
689 *Neuron Disord* 2000;1(5):293-9.
690 43 Igaz LM, Kwong LK, Lee EB, et al. Dysregulation of the ALS-associated gene TDP-
691 43 leads to neuronal death and degeneration in mice. *J Clin Invest*
692 2011;121(2):726-38.
693 44 Walker AK, Spiller KJ, Ge G, et al. Functional recovery in new mouse models of
694 ALS/FTLD after clearance of pathological cytoplasmic TDP-43. *Acta*
695 *neuropathologica* 2015;130(5):643-60.

45 DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 2011;43(5):491-8.
46 McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
47 Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;11(1110):11.10.1-11.10.33.
48 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 2010;38(16):e164.
49 Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536(7616):285-91.
50 Kann O, Kovacs R. Mitochondria and neuronal activity. *American journal of physiology Cell physiology* 2007;292(2):C641-57.
51 Lehmer C, Schludi MH, Ransom L, et al. A novel CHCHD10 mutation implicates a Mia40-dependent mitochondrial import deficit in ALS. *EMBO molecular medicine* 2018;10(6):e8558.
52 Straub IR, Janer A, Weraarpachai W, et al. Loss of CHCHD10-CHCHD2 complexes required for respiration underlies the pathogenicity of a CHCHD10 mutation in ALS. *Human molecular genetics* 2018;27(1):178-89.
53 Ronchi D, Riboldi G, Del Bo R, et al. CHCHD10 mutations in Italian patients with sporadic amyotrophic lateral sclerosis. *Brain* 2015;138(Pt 8):e372.
54 Dobson-Stone C, Shaw AD, Hallupp M, et al. Is CHCHD10 Pro34Ser pathogenic for frontotemporal dementia and amyotrophic lateral sclerosis? *Brain* 2015;138(Pt 10):e385.

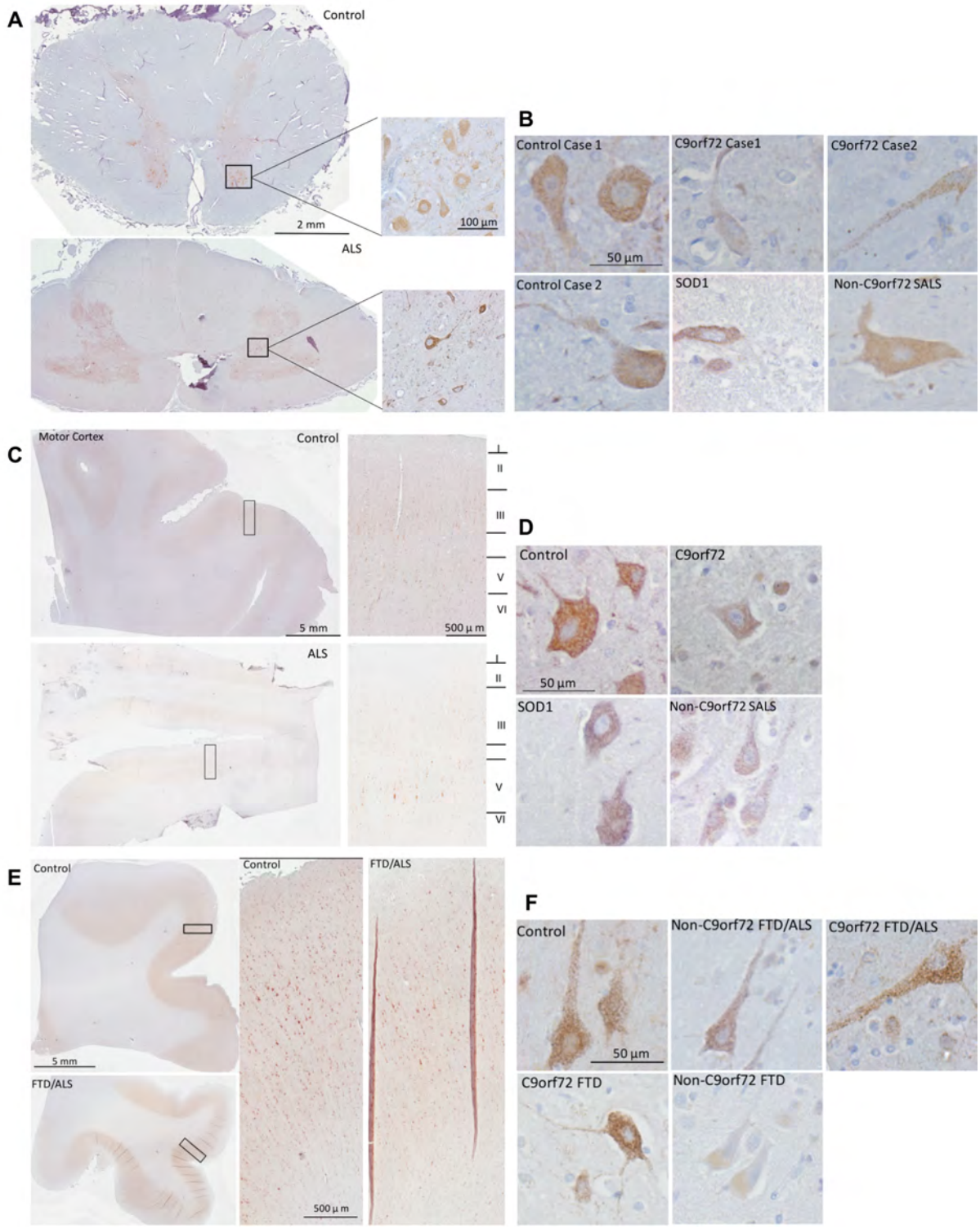


Figure 1 IHC staining of CHCHD10 in (A, B) spinal cord, (C, D) motor cortex and (E, F) frontal cortex in human post-mortem tissues. A, C, E are ScanScope images of the whole section from one control and one patient from each location, and the zoomed-in view of the boxed area. B, C, D were neurons from control and patients with different genotypes taken with Zeiss Axio Imager using a 20x lens. A, C, E showed that in all three locations, CHCHD10 showed positive staining primarily in

the grey matter, and strong immunoreactivity specifically to the neurons. Reduced numbers of CHCHD10 positive neurons were seen in all three locations (A, C, E, Boxed area). In all three locations, CHCHD10 showed primarily cytoplasmic localisation. No significant difference was seen in terms of subcellular location between control and patients, and between patients with different genotypes (B, D, E).

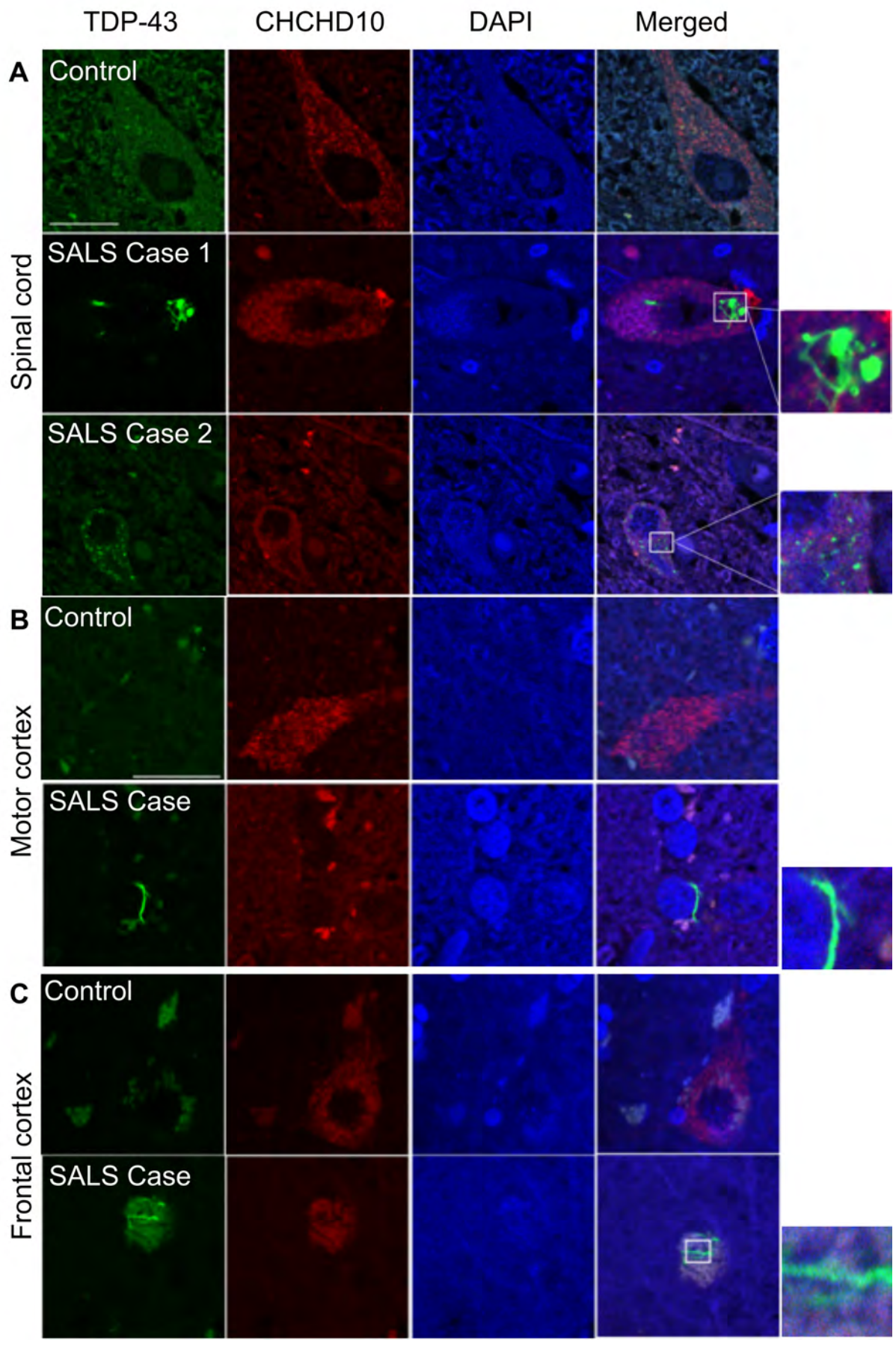

TDP-43
CHCHD10
DAPI
Merged
A
Control
Spinal cord
SALS Case 1
SALS Case 2
B
Control
Motor cortex
SALS Case
C
Control
Frontal cortex
SALS Case

Figure 2 CHCHD10 did not colocalise with TDP-43 inclusions in spinal cord (A), motor cortex (B) or frontal cortex (C) tissues. TDP-43 inclusions were labelled with an antibody that is specific to phosphorylated TDP-43 (green) and CHCHD10 was labelled with anti-CHCHD10 antibody (red). Colocalisation between TDP-43 inclusions and CHCHD10 was not observed in most of the neurons, except for a small number of motor cortex neurons, where partial colocalisation was evident (B, insets). Images were photographed using a 63X lens. Scale bar: 20μm.

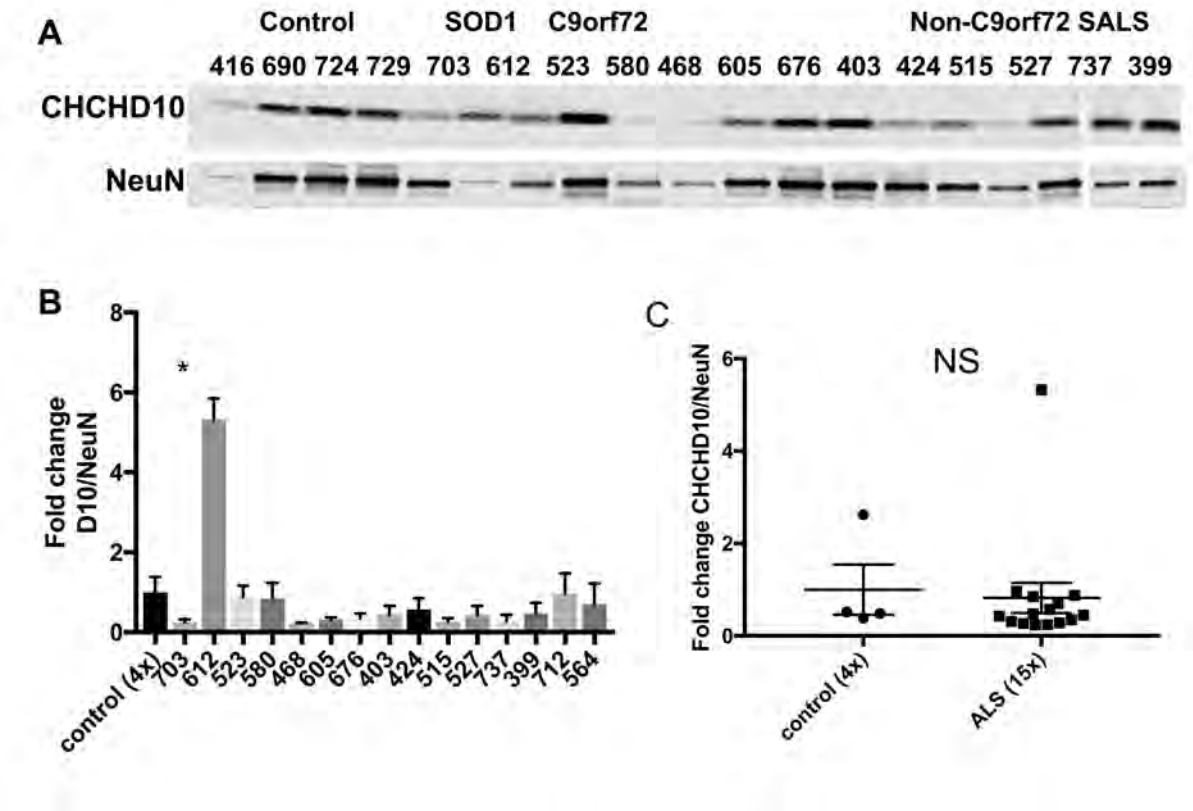


Figure 3 CHCHD10 showed variable expression in post-mortem motor cortex tissues. (A) Western blotting of control or ALS motor cortex tissues; (B) Semi-quantification of CHCHD10 expression normalised NeuN showed variable expression between individual cases. (C) Semi-quantification data of B was grouped as control and ALS. A decreasing trend was seen in ALS cases, although the difference is not significant between control and patient.

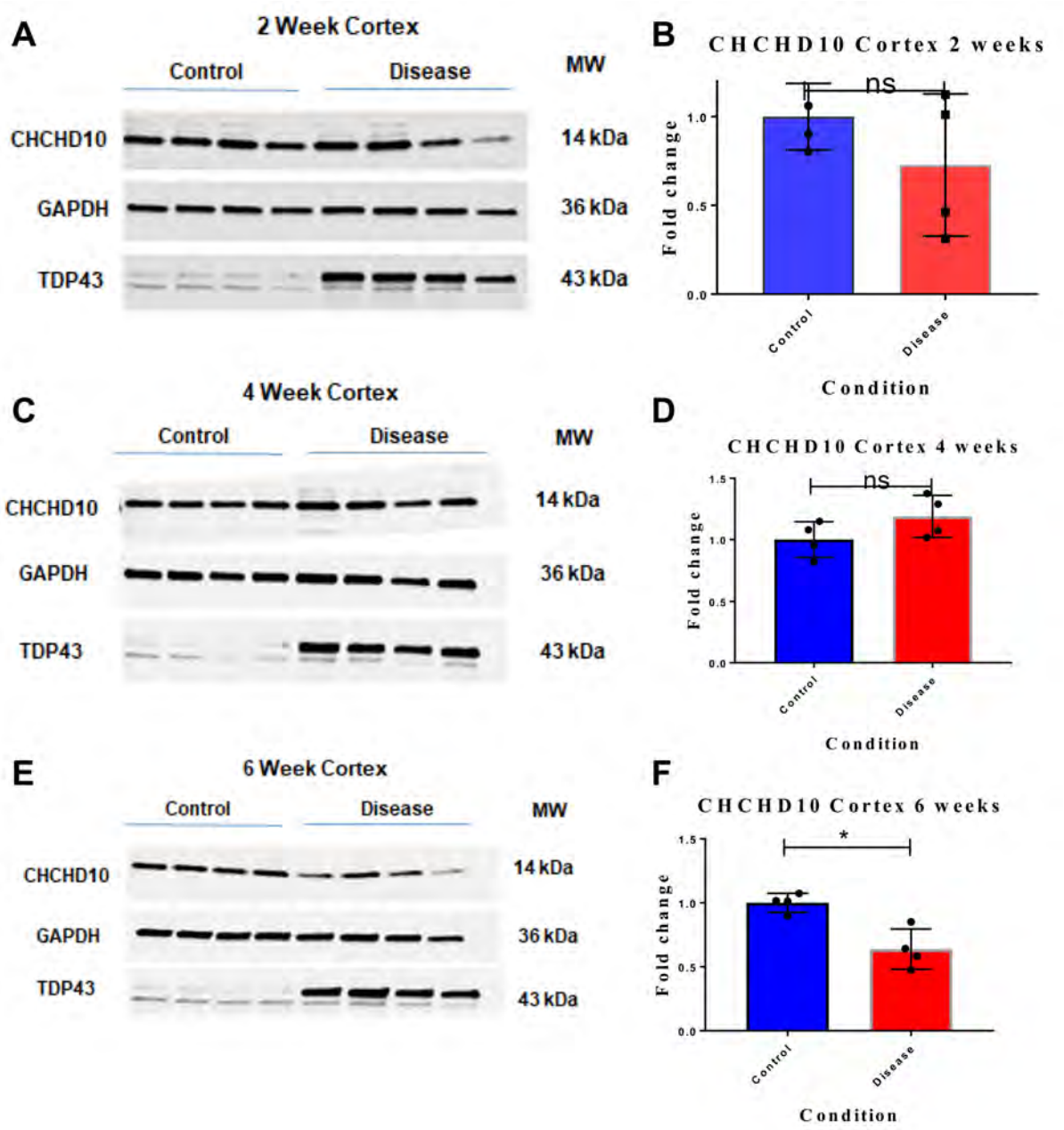
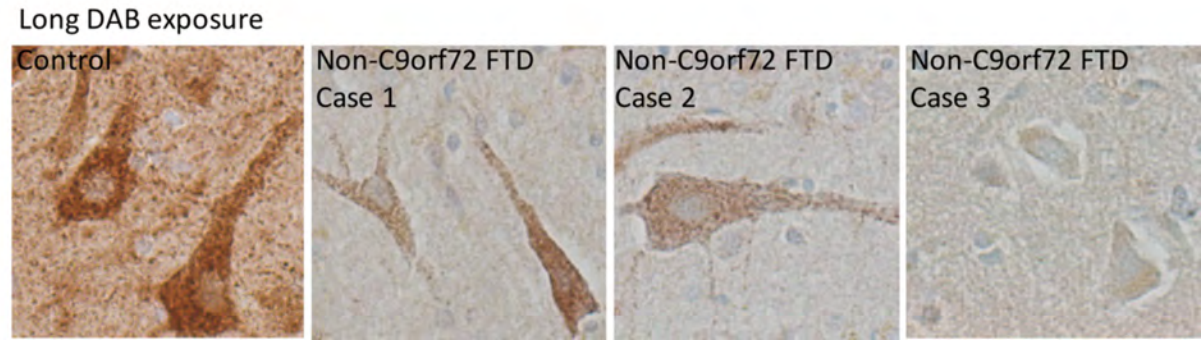
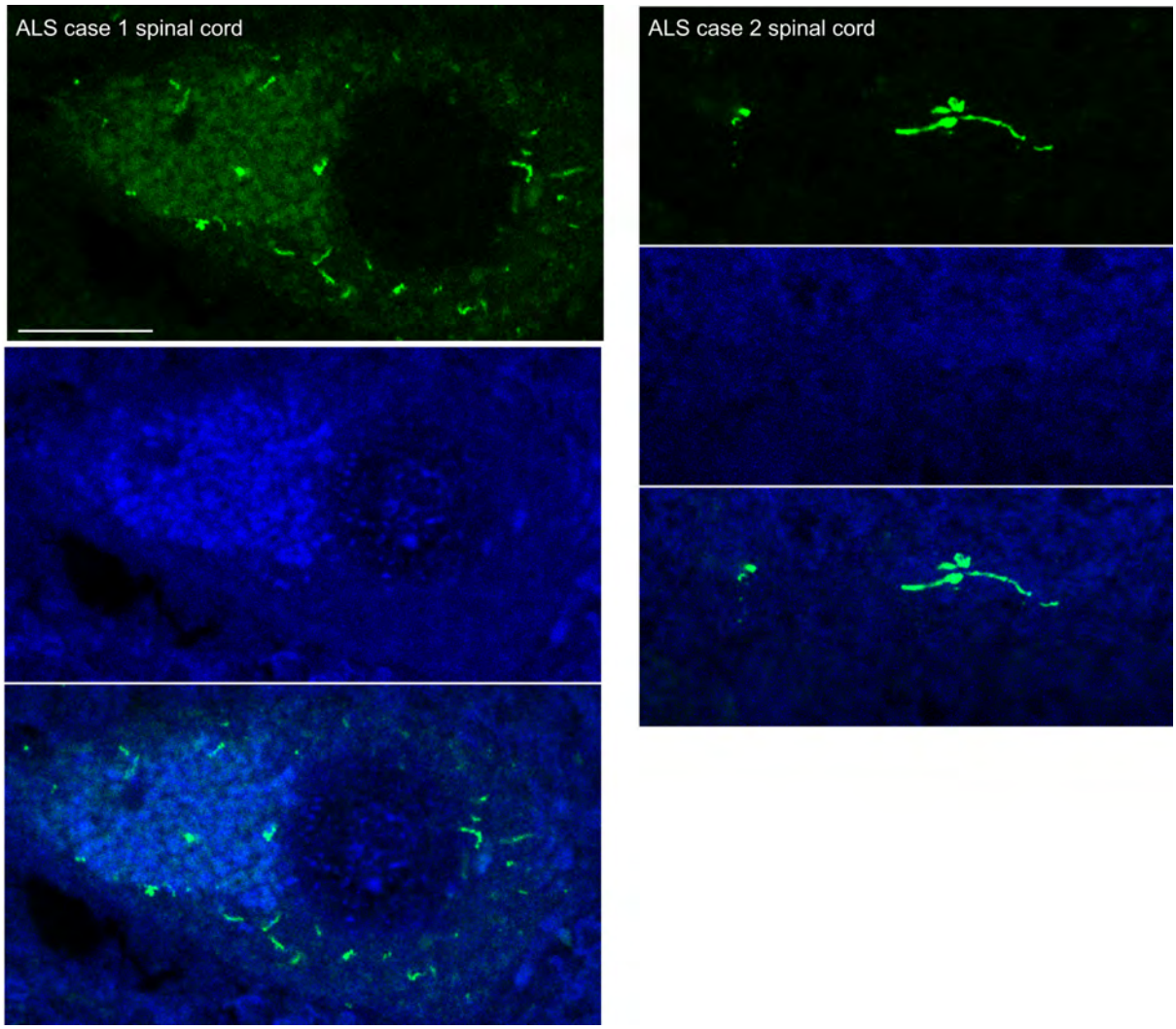


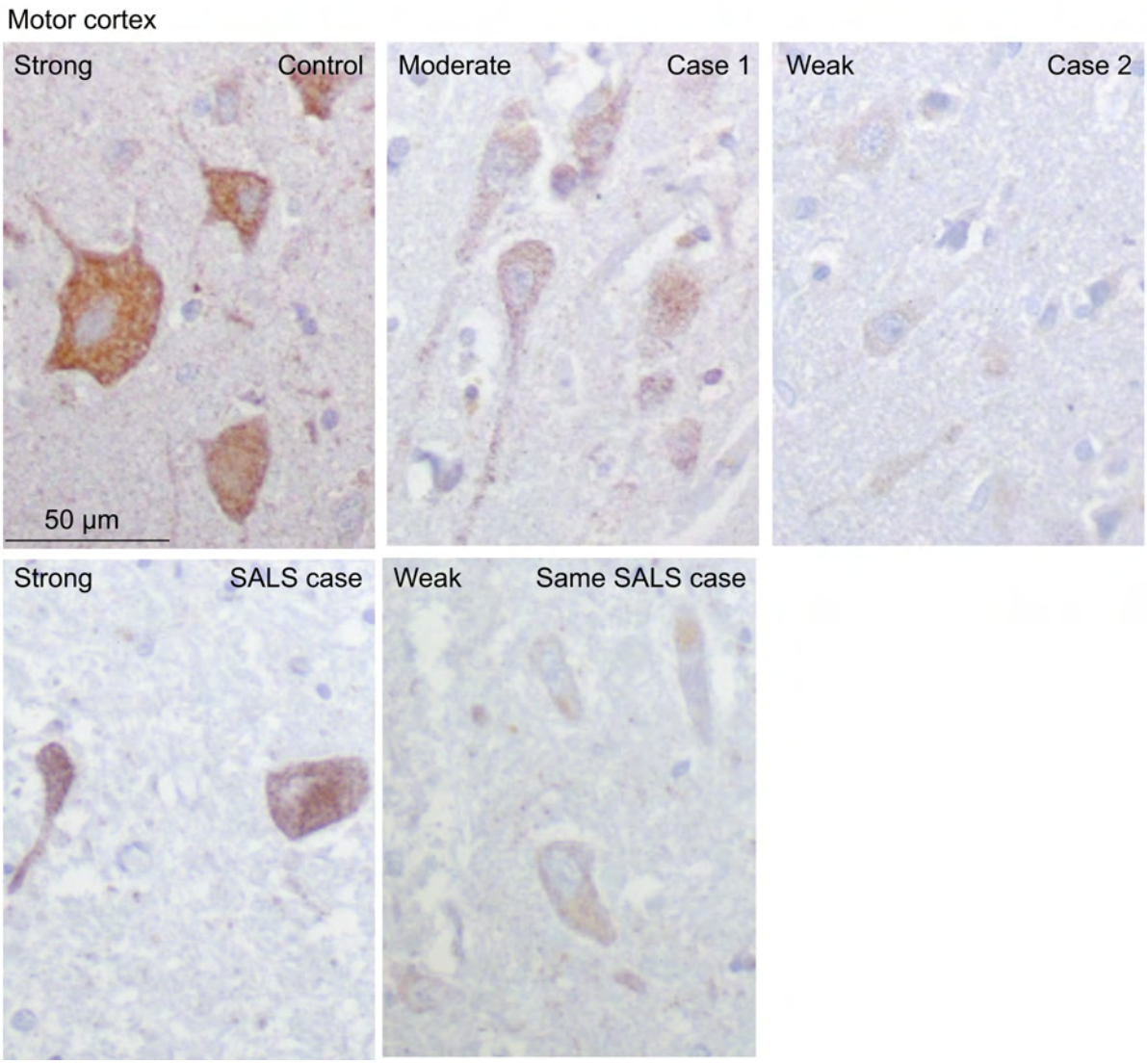
Figure 4 CHCHD10 expression in a mouse model of ALS (rNLS TDP-43 mice). Mouse brain (cortex) was collected at at pre-symptomatic (2 weeks) and symptomatic (4 & 6 weeks) TDP-43 mice and litter-matched controls (n=4/group) and was immunoblotted with CHCHD10 antibody. No significant changes in the expression of CHCHD10 was seen at 2 and 4 weeks post-disease onset. In contrst, a significant decrease in the expression of the CHCHD10 gene was observed at 6 week post-onset in diseased mice compared to controls.



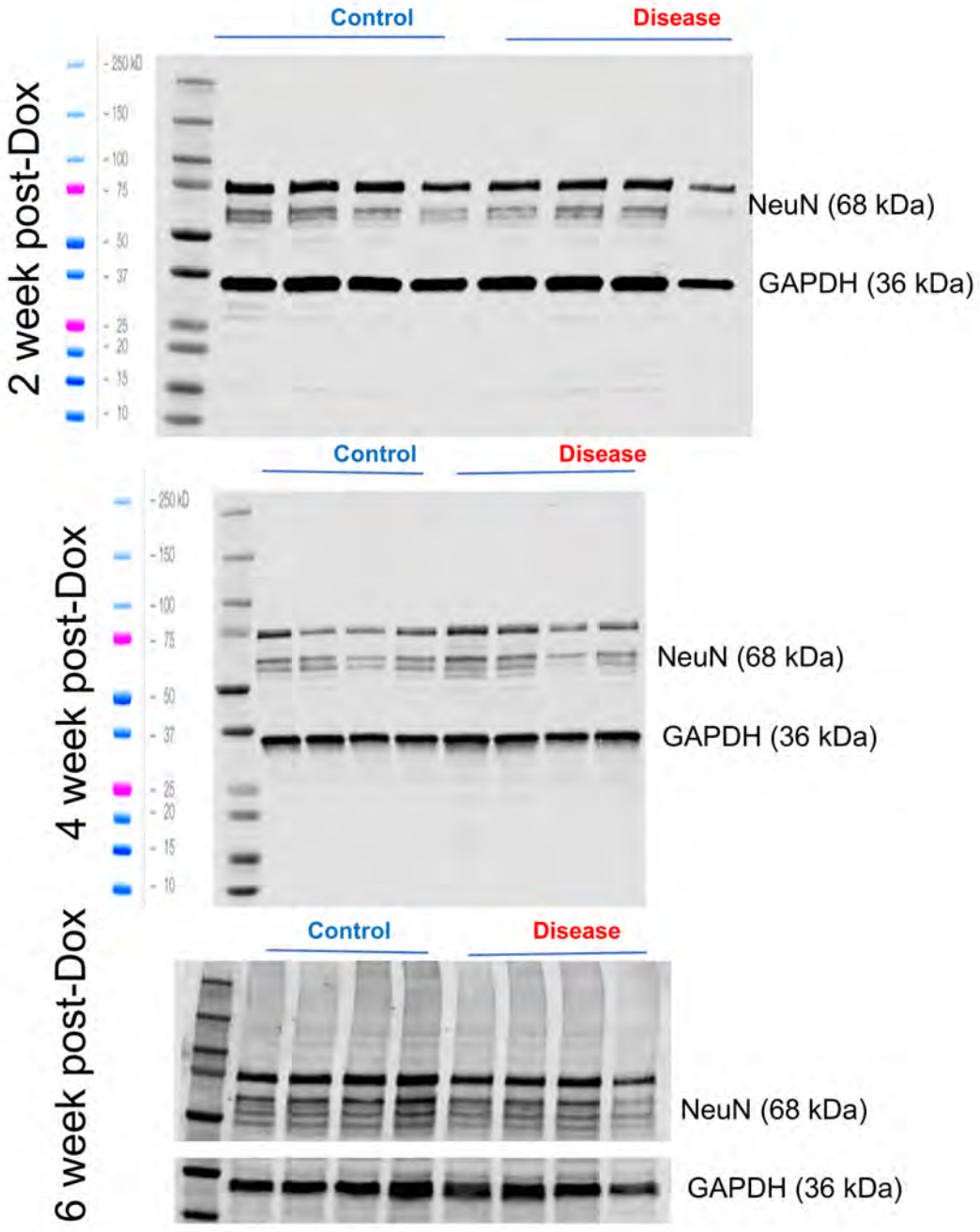
Supplementary figure 1 IHC staining of CHCHD10 in Non-C9orf72 FTD frontal cortex tissues with long DAB exposure. Frontal cortex sections from three Non-C9orf72 FTD cases were incubated with DAB for 5 min. While Case 1 and 2 showed increased staining, Case 3 remained weak.



Supplementary figure 2 CHCHD10 (green) formed dense dot inclusion-like structures in a few neurons from two SALS spinal cord cases and one SALS motor cortex cases.



Supplementary figure 3 Examples of variable IHC staining levels within the same case and in between cases



Supplementary figure 4 NeuN staining showed no significant difference between control and diseased mice at 2, 4 or 6 weeks off-Dox.

Table S1 *CHCHD10* varaints identified in Australian ALS and FTD, and associated allele frequenes in cases and controls

					GnomAD NFE		DMCC		MGRB	
Variant	**Function**	**dbSNP ID**	**Cohort**	**AAF**	**AAF**	**p-value**	**AAF**	**p-value**	**AAF**	**p-value**
c.*65C>T	UTR3	.	SALS	0.0008	0	0.076	*na*	*na*	0.0004	1
c.*53C>T	UTR3	rs113889670	SALS	0.0008	0.0001	0.209	*na*	*na*	0	0.3540
c.*8G>A	UTR3	rs372342375	SALS	0.0008	0.0005	0.439	*na*	*na*	0.0009	1
c.409+27C>G	intronic	rs140182	SALS	0.8040	0.8292	0.484	0.8410	0.0073	0.8360	0.0185
			FALS	0.7570	0.8292	0.493	0.8410	0.4390	0.8360	0.4802
.410C>T, p.Tyr104Tyr	synonymous SNV	rs80027270	SALS	0.8050	0.8280	0.524	0.8400	0.0121	0.8350	0.0267
			FALS	0.7320	0.8280	0.363	0.8400	0.3143	0.8350	0.3499
c.262-57C>T	intronic	rs755979336	SALS	0.0016	0.0010	0.371	*na*	*na*	0.0026	0.7202
c.262-149G>A	intronic	rs149955095	SALS	0.0199	0.0142	0.119	*na*	*na*	0.0157	0.3487
c.262-239C>G	intronic	rs9608181	SALS	0.0901	0.0888	0.877	*na*	*na*	0.0804	0.3417
c.262-294T>C	intronic	rs131441	SALS	0.8054	0.8250	0.598	*na*	*na*	0.8344	0.0337
c.262-341G>C	intronic	rs6003876	SALS	0.0199	0.0201	1	*na*	*na*	0.0197	1
c.262-344T>A	intronic	rs572584379	SALS	0.0024	0.0030	1	*na*	*na*	0.0026	1
c.262-515T>C	intronic	rs131442	SALS	0.8118	0.8253	0.710	*na*	*na*	0.8383	0.0500
c.261+276G>A	intronic	rs73396548	SALS	0.1396	0.1425	0.835	*na*	*na*	0.1425	0.8402
c.261+152G>A	intronic	rs73396549	SALS	0.1380	0.1411	0.834	*na*	*na*	0.1403	0.8791
c.261+135A>G	intronic	rs73158728	SALS	0.0016	0.0038	0.325	*na*	*na*	0.0026	0.7202
c.261+99A>G	intronic	rs131443	SALS	0.4880	0.5119	0.360	*na*	*na*	0.5232	0.0490
c.261+73G>A	intronic	rs80167838	SALS	0.0207	0.0227	0.766	*na*	*na*	0.0166	0.4288
c.261+11A>G	intronic	rs131444	SALS	0.8840	0.9040	0.303	0.88	0.7806	0.8960	0.2819
			FALS	0.8660	0.9040	0.752	0.88	0.5411	0.8960	0.7949
.234G>A, p.Ser78Ser	synonymous SNV	rs111527940	FALS	0.0139	0.0111	0.076	0.0150	1	*na*	*na*
c.48A>C, p.Pro16Pro	synonymous SNV	rs179468	FALS	0.7500	0.8906	0.136	*na*	*na*	0.8911	0.9093
c.20-42T>C	intronic	rs1023954590	SALS	0.0008	0.00006	0.146	*na*	*na*	0	0.3540
c.42-118G>C	intronic	rs113097524	SALS	0.0008	0	0.076	*na*	*na*	0.0022	0.4330
c.41+46T>G	intronic	rs62241575	SALS	0.0718	0.0483	0.0003	0.0850	0.5060	0.0752	0.7375
			FALS	0.0130	0.0483	0.167	0.0850	0.0151	0.0752	0.1624
c.41+7C>T	intronic	rs141526972	SALS	0.0016	0.0022	0.237	*na*	*na*	0.0017	1
			FALS	0.0208	0.0484	0.169	*na*	*na*	0.0017	0.2613

AFF, alternate allele frequency; NFE, Non-Finnish European; DMCC, Diamantia control cohort; MGRB, Medical Genomic reference bank

4.3.3 Co-authored publications

Throughout the course of this project, the candidate was approached by colleagues and national and international collaborators to assist in the replication phase of SNP association analyses. To achieve this, high-throughput custom TaqMan genotyping of the SNP under investigation was performed through a large cohort of Australian SALS patients and non-related control individuals. Table 4.1 summarises these analyses, and provides the reference for the resulting publication co-authored by the candidate.

Table 4.1: **Co-authored publications resulting from ALS gene screening of Australian cohorts.**

Gene	Risk SNP	Analysis	Summary	Publication
C21orf2 (and *MOBP*)	rs7508772 and rs616147	Custom TaqMan genotyping for 774 SALS & 785 non-related control individuals	Van Rheenen *et al.* used imputation and mixed-model association analysis to identify *C21orf2* as an ALS risk gene, and additionally identifued *MOBP* and *SCFD1* as newly associated risk loci. As part of the replication phase of this study, the candidate performed custom TaqMan genotyping for the candidate ALS risk SNPs, rs75087725 (*C21orf2*) and rs616147 (*MOBP*) in an Australian cohort of SALS and non-related control individuals. Fisher's exact testing suggested there was no association between either SNP and SALS risk in our cohort. However, when tested using logistic regression and combined as part of the larger international replication cohort in a meta-analysis, our collaborators found that the significant association between each of these SNPs and ALS was replicated. **The overall finding of this study was the identification of *C21orf2* as an ALS risk gene.**	Paper A1: Appendix A.5.1
GPX3-TNIP1	rs10463311 (rs4958872 as proxy) and rs9906189	Custom TaqMan genotyping for 431 SALS & 567 non-related control individuals	Benyamin *et al.* conducted a cross-ethnic meta-analysis genome-wide association study (GWAS) in SALS, and identified the *GPX3-TNIP1* locus to be significantly associated with disease. A gene-based analysis also implicated *GGNBP2* as being associated with disease. As part of the replication phase of this study, the candidate performed custom TaqMan genotyping. This was performed for rs4958872 (*TNIP1*, which was used as a proxy for the SNP of interest, rs10463311 (*GPX3*, LD r^2=1)) and rs9906189 (*GGNBP2*). In our cohort, neither genotyped SNP was associated with disease, however when combining our proxy data with that from another Australian cohort, Benyamin *et al.* found rs10463311 to be significantly associated with SALS. **The overall finding of this study was the identification of rs10463311 as an ALS risk allele.**	Paper A2: Appendix A.5.2
HNRNP genes	rs2588882	Custom TaqMan genotyping for 160 SALS & 115 non-related control individuals	Our group set out to investigate the contribution of the *HNRNP* gene family to the genetic and pathological basis of ALS. As part of this study, *HNRNP* genes were screened through FALS patients by the candidate, however no novel mutations were identified. Another aspect of this study found that two *HNRNPA3* SNPs, rs2588882 and rs8470 were over represented in FALS probands compared to three separate control cohorts. As part of the replication phase, the candidate performed custom TaqMan genotyping for rs2588882 in SALS and non-related controls, the results of which suggested the association was not present in SALS patients. **The overall finding of this study was the rarity of *HNRNP* gene mutations in Australian ALS, and the description of a unique hnRNPA3 related protein pathology in *C9orf72* expansion positive ALS patients.**	Paper A3: Appendix A.5.3

4.4 Discussion

ALS is a complex disorder and is highly heterogenous, both genetically and phenotypically. In order to understand how and why disease develops, it is imperative that the scope of this heterogeneity is accurately characterised. Interestingly, this heterogeneity can be partly explained by the known ALS genes and mutations, where distinct frequency patterns and phenotypic correlations have emerged. For instance, the frequency with which ALS mutations cause disease can vary significantly between populations, while particular ALS mutations seem to predispose patients to a certain phenotypic pattern of disease. This Chapter has presented data that significantly adds to the scope of our current knowledge of the ALS gene mutation spectrum. The unique pattern of ALS gene frequencies observed among Australian patients for established ALS genes (Paper I; Section 4.3.1) and the recently reported *CHCHD10* gene (Manuscript II; Section 4.3.2) have been described. The scope of ALS risk variants has also been established by analysis of the genes *CHCHD10* (Manuscript II; Section 4.3.2), *C21orf2* (Paper A1; Appendix A.5.1), *MOBP* (Paper A1; Appendix A.5.1), *GPX3-TNIP1* (Paper A2; Appendix A.5.2) and *hnRNPA3* (Paper A3; Appendix A.5.3). Additionally, further characterisation of the ALS genes has been provided by description of their correlations with clinical characteristics (Paper I; Section 4.3.1), as well as the pathology of the CHCHD10 protein in patient neuronal tissue (Manuscript II; Section 4.3.2).

Though the Australian population are predominately of European ancestry, a combination of factors such as geographical isolation coupled with migration patterns and multiculturalism result in a unique genetic background for this population. As such, it is to be expected that a distinct spectrum of ALS gene mutation frequencies and associations would be present among Australian patients. While the mutational frequencies of the established ALS genes in the Australian FALS cohort were found to be similar to that seen in European-based populations, it is interesting to note that the *SOD1* p.A5V mutation, the most common single point mutation causing ALS in the North American population (Andersen, 2006), was distinctly absent in our Australian FALS cohort. Similarly, while mutations in *CHCHD10* have consistently been identified in European populations (Bannwarth et al., 2014; Dols-Icardo et al., 2015; Kurzwelly et al., 2015; Muller et al., 2014; Perrone et al., 2017), no such mutations were present in Australian patients. This likely reflects the pattern of migration of Americans and Europeans to Australia, and is an important consideration for genetic analysis and screening prioritisations. As such, using the Australian cohort, there

is mixed ability to replicate international reports of novel causal or associated ALS genes. To confirm the ancestry of the Australian patient cohort, principal components analysis will be performed. Unfortunately, this is not yet possible as the computing capacity required to handle raw data files from the complete ALS patient cohort (FALS and SALS) is not yet available. When completed, this will shed light on the extent to which genetic diversity in the Australian population differs to that in European populations.

The benefits of our findings are multifaceted. From a medical research perspective, our gene frequency data, together with that from around the world, can be used to better inform downstream study designs. For instance, when choosing an ALS mutation as the basis of a novel animal model, one which has been shown to segregate with disease in multiple families should be chosen in favour of another found to be present in a single proband patient. This is vital to ensure disease models are based on mutations with infallible support for pathogenicity, as well as to make these models relevant for larger numbers of patients. Further, biomarker or therapeutic studies aimed at mutations that are rare amongst Australian patients may be better suited for trial in another population. Such genetically informed decisions will ensure that the most applicable and relevant research studies and clinical trials are performed.

Clinically, our frequency data can be used to prioritise diagnostic gene screening efforts for Australian ALS patients, which may potentially influence downstream carrier and/or preimplantation embryonic screening choices. Interestingly, preliminary data suggest great promise for therapeutic treatments based on genetic predisposition to disease. A clinical trial for CuATSM in *SOD1* patients is underway after having shown positive effects in mouse models (Hilton et al., 2017; Williams et al., 2016a). Another clinical trial for lithium carbonate has shown that those patients who carry an ALS associated SNP in *UNC13A* show increased survival with treatment, compared with non-SNP carrier patients (van Eijk et al., 2017). The ways in which particular ALS mutations correlate with age of onset or disease duration may also have utility in the clinic. An "at-risk" individual may be influenced by the knowledge that their family mutation associates with early or late disease onset, and as a result opt in or out of genetic testing. Further, if a patient can be given an estimated disease duration, they may be able to make more appropriate and timely decisions about symptom management and quality of life strategies.

The association of known population-based SNPs to disease is becoming increasingly important in our knowledge of the factors that contribute to disease risk. The high rate of sporadic ALS, and the late onset of disease suggest that there may be an accumulation of risk factors contributing to the eventual onset of disease (discussed Chapter 8, Section 8.2.1.6). Identifying known population-based SNPs that increase the risk of developing disease may become a potentially important ALS risk assessment tool in the future. Indeed, association testing was critical to the discovery of the pathogenic expansion in *C9orf72* (see Section 1.4.1.4). Interestingly, Jones et al. (2013) found that one SNP associated with this pathogenic expansion is also associated with disease in ALS patients negative for the expansion. Future investigations in our expanded cohort of Australian FALS and SALS patients are planned to determine whether such an association is reproducible.

As a rare disease, a global effort is required for an accurate characterisation of ALS genetics. This rarity coupled with the small population size of Australia, causes innate difficulties in collecting patient sample cohorts of adequate size to perform accurate assessments of mutation frequencies and associations. Fortunately, long running clinical collection programs and collaborations have allowed our laboratory to establish such patient cohorts, which has enabled these unique genetic investigations into Australian ALS. This has allowed the intricacies of the Australian spectrum of ALS genes to be better understood, and has also provided crucial insights to population-specific disease associations.

"I don't need sleep, I need answers"

Sheldon Cooper - The Big Bang Theory

5 Investigation of candidate ALS genes

5.1 Introduction

This Chapter addresses the second part of Aim 2 of this thesis; to investigate candidate ALS genes in familial and sporadic Australian ALS patients to identify novel or known ALS mutations and/or associated genetic variants. Its purpose is to determine whether known and candidate ALS genes contribute to the cause of ALS among Australian patients.

As was established in Chapter 1, Section 1.4, ALS is an exceptionally genetically heterogeneous disease, with at least 25 causal genes, and a further 12 disease associated genes identified to-date. Paper I (Chapter 4, Section 4.3.1) showed that the genetic landscape of Australian FALS is unique, with 21 distinct mutations in eight different genes causing disease in this patient population. A noteworthy conclusion of this paper was that almost 40% of Australian FALS patients did not carry mutations in the known ALS genes. Unfortunately, most ALS families for whom a mutation remains to be identified, only have DNA available from the proband, and therefore family-based linkage or segregation analysis is not possible. However, a strong family history of disease in these patients suggests that they almost certainly carry a novel, rare genetic mutation that causes ALS. As such, alternate strategies are necessary to

identify the underlying causal ALS mutation in these FALS patients.

Candidate gene screening strategies have had great success in ALS research. Drawing on the results of genetic linkage analysis, candidate gene screening approaches successfully identified the ALS genes *FUS* (Vance et al., 2009) and *UBQLN2* (Deng et al., 2011), among many others (see Chapter 1, Section 1.4.1.6). Further, a number of ALS genes including, *FIG4* (Chow et al., 2009), *SQSTM1* (Fecto et al., 2011), *GLE1* (Kaneb et al., 2015), and most notably *TARDBP* (Sreedharan et al., 2008), were investigated as candidate genes in ALS families owing to functional evidence suggesting that their encoded protein product was involved in ALS pathogenesis.

There is also a substantial body of evidence supporting a genetic underpinning to sporadic disease (see Chapter 1, Section 1.4.2). As described, multiple genetic risk factors have recently been reported in SALS patient cohorts. In addition, a small proportion of SALS patients are likely to be misclassified FALS patients, for whom limited family histories are available. Indeed, some SALS patients do carry known ALS gene mutations, such as those in *CCNF*, *TARDBP*, *FUS*, *EWSR1* and *C9orf72* (Couthouis et al., 2012; Sreedharan et al., 2008; Vance et al., 2009; Williams et al., 2013, 2016b). Though likely rare, novel causal gene mutations in (apparently) sporadic patients are difficult to identify, and their identification will require alternate strategies, such as candidate gene screening.

Here, a candidate gene approach was employed to identify novel genetic contributors to ALS pathogenesis in Australian patients with no known causal gene mutation. The candidate genes analysed in this Chapter possess evidence suggestive of a role in ALS from a range of different research strategies including genetics, proteomics and animal models. The majority of these candidate gene analyses were conducted using FALS patients with an unknown ALS gene mutation, most of whom were probands. Some candidate genes were also screened through a large cohort of SALS patients, which became available in the later stages of this candidature. Two types of genetic variants were targeted in this Chapter. The first being novel non-synonymous variants potentially causing ALS (candidate mutations). Secondly, known population-based SNPs (both rare or common) found in healthy individuals were investigated for their potential to confer an increased disease-risk or protection against disease, based on their over- or under- representation in ALS patients, respectively.

5.2 Subjects and methods

5.2.1 Subjects

Datasets from two patient cohorts were analysed in this Chapter. The first consisted of whole-exome sequencing (WES) data from 81 Australian FALS affected individual (including 61 probands) from 69 families, with an unidentified ALS causal mutation. These affected individuals were previously screened for known ALS genes in Chapter 4, Paper I (Section 4.3.1) and Manuscript II (Section 4.3.2). Whole-genome sequencing (WGS) data from 635 Australian SALS affected individuals negative for the *C9orf72* hexanucleotide repeat expansion made up the second cohort. Further details of both cohorts are provided in Chapter 2, Section 2.1.

5.2.2 Pipeline for screening candidate ALS genes and association analysis

Figure 5.1 describes the bioinformatics pipeline applied to FALS patient WES data and SALS patient WGS data to identify variants found within a given candidate gene, and to determine whether any novel causal gene mutations or disease-associated population-based SNPs were present in either patient cohort. The custom bioinformatics scripts applied were either developed for general NGS-based genetic analysis as part of Chapter 3, or specifically for candidate gene analysis in this Chapter, as detailed below.

5.2.2.1 Candidate gene screening in FALS patients with an unidentified ALS mutation

Firstly, to facilitate association analysis, control allele count data from the ExAC, DACC and MGRB databases (as described in Chapter 2, Table 2.4) was first appended to the 137-sample WES VCF (containing data from all FALS affected individuals, described in Chapter 2, Table 2.1) using the Custom Scripting strategy developed in Chapter 3, Section 3.5.3. The Custom R markdown Script in Appendix A.2.4 was developed to perform candidate gene analysis separately for the different cohorts of FALS affected individuals present within this VCF, particularly the 81 FALS affected individuals with an unidentified ALS mutation.

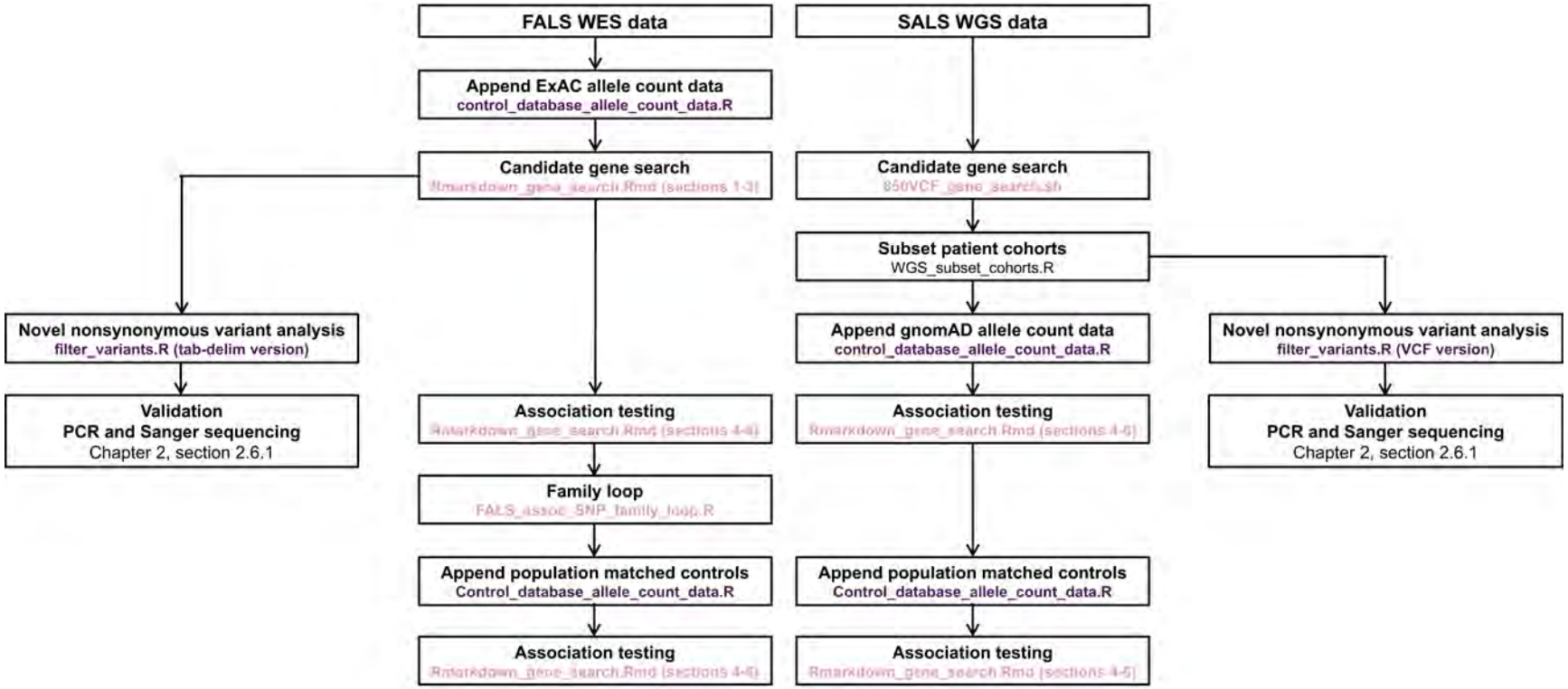


FIGURE 5.1: **Candidate ALS gene analysis workflow.** WES data from 81 FALS patients with an unidentified ALS causal mutation and WGS data from 635 SALS patients were bioinformatically interrogated to identify variants present in a given candidate ALS gene. Analysis included the identification of novel non-synonymous variants with the potential to cause disease, and association analysis to identify known population-based SNPs over- or under-represented in ALS patients. Each analysis step was completed using custom bioinformatics scripts. Those shown in purple were developed for general NGS-based genetic analysis as part of Chapter 3, while those in pink were specifically developed for use in the current Chapter.

Following initial import of the complete 137-sample WES VCF, FALS patient cohorts were subsetted to individual data frames using a *subset* command. The ANNOVAR annotation column, "Gene.refGene" was then parsed for a given list of candidate genes using another *subset* command and the value matching operator, *%in%*. This facilitated extraction of all variants present in the candidate gene(s), resulting in a data frame consisting of candidate gene(s) variants and their associated meta and INFO information, as well as sample information for each member of a given cohort. A series of arithmetic functions were then applied to the data frame in order to add new columns containing affected individual genotype counts and allele counts. For each candidate variant with control allele count data available, an R *for-loop* function was used to perform Fisher's exact tests to compare allele counts from affected individuals with those from the appended control database.

Novel non-synonymous variant analysis

To be considered novel, a variant was required to be present in two or less individuals from the Non-Finnish European (NFE) cohort from the ExAC (n=60,706 total individuals; n=33,370 NFE individuals) and gnomAD (n=138,632 total individuals; n=63,369 NFE individuals) control databases, as well as in the Australian control databases, DACC (n=967) and MGRB (n=1,144). Variants present in one or two individuals from a control database were also retained in novel variant analysis as they were sufficiently rare to potentially represent a technical sequencing/bioinformatic error, or be present in an asymptomatic ALS patient (discussed in detail in Section 6.4.2.2). The majority of filtering was performed using the Custom Script 3.11. This included removing variants present in the ExAC, DACC and MGRB control databases by assessing the value of their respective allele count column. Similarly, only non-synonymous variants were retained in analysis by evaluating the annotated value for the "ExonicFunc.refGene" field using this same scripting strategy. Any remaining variants were screened through the gnomAD control database using the web browser interface (http://gnomad.broadinstitute.org/).

Where appropriate, visual inspection of the relevant pedigree structure was carried out to ensure the novel, non-synonymous candidate variant segregated with disease within the family. Each remaining candidate mutation was assessed for its potential ALS pathogenicity using the *in silico* pipeline developed in Chapter 6, Section 6.2.3. Additionally, the Project MinE web browser (http://databrowser.projectmine.com/), ALS data browser (ALSdb; http://alsdb.org/) and ALS variant server (AVS; http://als.umassmed.edu/)

were interrogated for each remaining candidate mutation. WGS data from 635 SALS affected individuals (within the 850-sample VCF) and WES/WGS data for 247 FALS affected individuals from dbGAP (database of Genotypes and Phenotypes; https://www.ncbi.nlm.nih.gov/gap/; dbGaP Study Accession: phs000101.v5.p1) were also screened for each remaining candidate mutation using Custom Scripts (as per Appendix A.2.6, and a variation of line 54 of Appendix A.2.4, respectively).

Association testing

To assess whether any known population-based SNP variants in candidate ALS genes were associated with disease, scripting strategies were developed to compare allele counts between affected individuals and controls using a Fisher's Exact test. In the first instance, FALS affected individual allele counts were compared with those from the ExAC control database (n=60,706 exomes). Any variants with a p-value<0.05 were considered to be nominally significantly associated with disease following this analysis.

This FALS dataset included eight families with multiple affected members (n=20 individuals). To account for family-biased allele counts when determining associations, a complex Custom Scripting strategy was developed, to carry out association testing on each possible combination of affected individuals where just one member of each family was included (Appendix A.2.5). This script tested all 1,152 possible such combinations. Briefly, 1,152 data frames were first set up to contain sample information for all 1,152 affected individual combinations. Fisher's exact tests were then performed using affected individual allele counts calculated across each data frame. The results were then output to a separate results data frame in which p-values from each of the 1,152 combinations were presented, and subsequently visually assessed. A p-value<0.05 was required from each of the 1,152 combinations for the SNP to be considered nominally significantly associated with disease. SNP variants withstanding both the initial and family-loop analyses were then validated, by repeating this analysis using the Australian control cohorts, DACC (n=976 exomes) and MGRB (n=1,144 genomes), in place of those from ExAC. The 54 candidate genes screened using this strategy are listed in Table 5.1.

Following this baseline analysis, Bonferroni correction was applied to account for all 741 variants present in the 54 candidate genes. Therefore, the significance of each variant was reassessed following the about pipeline, though using a p-value threshold of 6.75×10^{-5}. For replication, Fisher's Exact testing was repeated for those variants

found to have nominally significant or Bonferroni-corrected significant association with disease, using the Project MiNE cohort of 4,366 SALS affected individuals and 1,832 control individuals.

5.2.2.2 Candidate gene screening in SALS patients

The 635 SALS affected individuals were among a total of 850 ALS and FTD affected individuals with data in the 850-sample WGS VCF (described in Chapter 2, Table 2.1). Therefore, initial candidate gene screening was performed on this 850-sample VCF in its entirety. This was achieved by developing the Custom bash Script A.2.6, which utilised the UNIX *awk* command, to search for a candidate gene name in the INFO column. The R script in Appendix A.2.7, was then applied to the resultant file to subset the SALS cohort from the resultant 850-sample VCF. Four candidate genes, *CHCHD2*, *CHCHD3*, *CHCHD6* and *TIA1*, were screened though SALS affected individuals using this pipeline.

Novel non-synonymous variant analysis

In order to determine whether any novel non-synonymous variants were present in a candidate gene among the 635 SALS affected individuals, the Custom R Script in Appendix A.2.8 was developed. As was completed for FALS analysis, each remaining candidate mutation was also assessed for its potential ALS pathogenicity using the *in silico* pipeline developed as part of Chapter 6, Section 6.2.3, and further screened through the additional ALS patient databases Project MinE (`http://databrowser.projectmine.com/`), the ALS data browser (ALSdb; `http://alsdb.org/`) and the ALS variant server (AVS: `http://als.umassmed.edu/`) using their respective web browsers. Additionally, WES/WGS data for 247 FALS affected individuals from dbGAP (`https://www.ncbi.nlm.nih.gov/gap/`; dbGaP Study Accession: phs000101.v5.p1) were also screened for each remaining candidate mutation using a Custom Script (as per line 54 of Appendix A.2.4).

Association testing

Association tests to compare allele counts between SALS affected individuals and controls were performed using a Custom R Script developed here (Appendix A.2.9). This included Fisher's exact tests for each variant identified in one of the four candidate genes using an R *for-loop*, which compared allele counts between SALS affected individuals and control individuals from either the complete gnomAD dataset

TABLE 5.1: **Candidate genes screened through FALS WES data.**

Candidate gene	Evidence justifying gene as an ALS candidate	Reference
PURA	Protein interacts with ALS mutated FUS proteins	Di Salvio et al. (2015)
C21orf2	ALS risk gene	van Rheenen et al. (2016)
MOBP	ALS associated gene	van Rheenen et al. (2016)
SCFD1	ALS associated gene	van Rheenen et al. (2016)
SPTBN4	Interrupted by transgene in motor impaired mouse model	Kichkin et al. (2017)
GLE1	Recently reported ALS gene	Kaneb et al. (2015)
MTHFSD	Differentially expressed in TDP-43 mouse model	MacNair et al. (2016)
DDX58	Differentially expressed in TDP-43 mouse model	MacNair et al. (2016)
CAMTA1	Associated with ALS patient survival	Fogh et al. (2016)
HNRNPA3	Related to the hnRNPA1 ALS gene	Kim et al. (2013)
EEF1A1	Collaborator proteomics work	unpublished data
EEF1A2	Collaborator proteomics work	unpublished data
EEF1A3	Collaborator proteomics work	unpublished data
EEF1B1	Collaborator proteomics work	unpublished data
EEF1B2	Collaborator proteomics work	unpublished data
EEF1B3	Collaborator proteomics work	unpublished data
EEF1B4	Collaborator proteomics work	unpublished data
EEF1D	Collaborator proteomics work	unpublished data
EEF1E1	Collaborator proteomics work	unpublished data
EEF1G	Collaborator proteomics work	unpublished data
NONO	Collaborator proteomics work	unpublished data
IKBKG	Collaborator proteomics work	unpublished data
ANXA11	Collaborator candidate FALS gene	unpublished data
ARPP11	Collaborator candidate FALS gene	unpublished data
GPX3	SNP associated with increased ALS risk	Benyamin et al. (2017)
TNIP1	SNP associated with increased ALS risk	Benyamin et al. (2017)
GGNBP2	SNP suggested as ALS risk gene	Benyamin et al. (2017)
ABCC2	Conference presentation suggested ALS risk gene	H. Kim, ASHG2016
TYBA4A	Conference presentation suggested differentially expressed gene	K. Belle, ASHG2016
UBA1	SMA gene	Ramser et al. (2008)
MTHFR	SNP suggested as ALS risk gene	Kuhnlein et al. (2011)
KIFAP3	SNP suggested as ALS modifier gene	Landers et al. (2009)
BICD2	SMA gene reported in juvenille ALS	Peeters et al. (2013) Neveling et al. (2013b) Oates et al. (2013)
CHCHD10	Recently reported ALS gene	Bannwarth et al. (2014)
CHCHD1	Gene family member of *CHCHD10*	N/A
CHCHD2	Gene family member of *CHCHD10*	N/A
CHCHD3	Gene family member of *CHCHD10*	N/A
CHCHD4	Gene family member of *CHCHD10*	N/A
CHCHD5	Gene family member of *CHCHD10*	N/A
CHCHD6	Gene family member of *CHCHD10*	N/A
CHCHD7	Gene family member of *CHCHD10*	N/A
PINK1	Parkinson's disease gene	Valente et al. (2004)
PARKIN	Parkinson's disease gene	Matsumine et al. (1997) Kitada et al. (1998)
CNR1	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
CNR2	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
FAAH	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
MGLL	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
DAGLA	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
DAGLB	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
NAPEPLD	Endocannabinoid system implicated in neurodegeneration	Pasquarelli et al. (2017)
KCND3	SNPs associated with PMA	unpublished data
CDT1	ALS gene substrate identified by collaborator	unpublished data
TIA1	Recently reported ALS gene	Mackenzie et al. (2017)
KIF5A	Recently reported ALS gene	Nicolas (2018)

Abbreviations: SMA, Spinal muscular atrophy; and PMA, Progressive muscular atrophy.

(n=123,136 exomes and 15,496 genomes), the gnomAD NFE subset (n=55,860 exomes and 7,509 genomes) or MGRB controls (n=1,144 genomes).

Following this baseline analysis, Bonferroni correction was applied to the p-value threshold of 0.05 for all 9,616 variants identified in WGS data across the four candidate genes screened through SALS. Re-analysis thus employed a significance threshold of $p<5.20\times10^{-6}$. For replication, Fisher's Exact testing was repeated for those variants found to have a nominally significant or Bonferroni-corrected significant association with disease, using the Project MiNE cohort of 4,366 SALS affected individuals and 1,832 control individuals.

5.3 Results

5.3.1 Novel non-synonymous candidate mutations

Table 5.2 provides a summary of the nine novel non-synonymous variants identified across the 54 candidate genes screened through 81 FALS affected individuals with an unidentified causal mutation, and the four candidate ALS genes screened through 635 SALS affected individuals. Of these nine candidate mutations, seven were identified in a single proband FALS patient; and therefore segregation with disease could not be established. Sanger sequencing confirmed all but one of these candidate mutations to be present in the respective patient DNA sample. The FALS affected individuals with candidate mutations in *EEFD1* and *DAGLB* underwent WES using whole-genome amplified (WGA) DNA samples, as insufficient quantities of non-amplified DNA were available for WES from these individuals. Sanger sequencing was performed on this WGA sample only for the *EEFD1* affected individual. An additional, non-amplified gDNA sample was available for the *DAGLB* affected individual, therefore Sanger sequencing was performed for both the amplified and non-amplified affected individual DNA samples. Sanger sequencing showed that while the *DAGLB* candidate mutation was present in the WGA DNA sample, it was in fact absent from the non-amplified gDNA sample. The two remaining candidate mutations were additional novel non-synonymous *TIA1* variants identified in one SALS affected individual each. Sanger sequencing validated both of these candidate SALS mutations within their respective patient DNA samples. Figure 5.2 shows an example chromatogram obtained from Sanger sequencing.

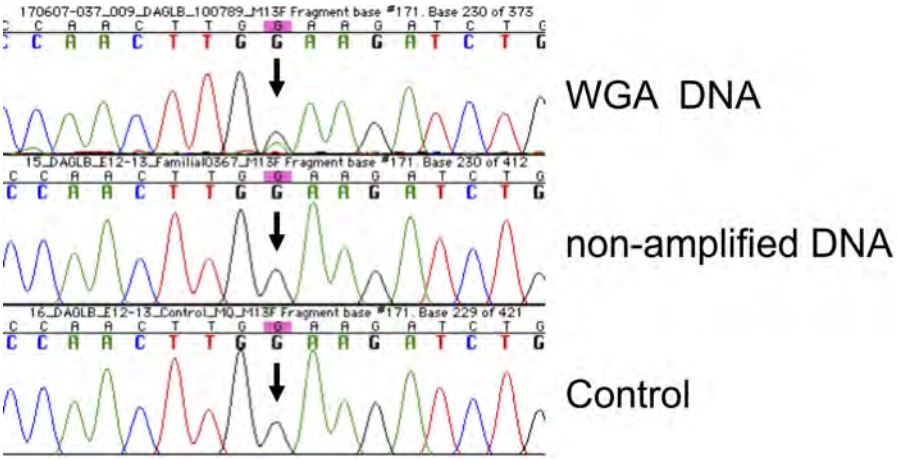


FIGURE 5.2: **Sequencing chromatogram for *DAGLB* candidate mutation.** The candidate mutation, *DAGLB* c.1516G>A; p.E506K, identified in a FALS proband patient from WES data was not validated by Sanger sequencing. The whole-genome amplified (WGA) DNA sample used for WES showed an inconclusive genotype upon Sanger sequencing validation, as seen in the sequencing chromatogram in the top panel. While a double peak is present, the wild-type allele peak height is consistent with that seen in the control sample (bottom panel), therefore this is an inconclusive genotype. However, the non-amplified DNA sample clearly has a single wild-type allele peak indicating a conclusive homozygous wild-type genotype (middle panel), which matches the control individual in the bottom panel.

In silico assessment of potential pathogenicity suggested that *DAGLB* p.E506K and *TIA1* candidate mutations p.A254G and p.P294L (identified in a FALS and SALS affected individual respectively) showed the most functional similarity to known ALS mutations, suggesting these were the more likely candidate mutations to be causing ALS in this cohort. The results of the *in silico* assessment of potential pathogenicity are presented in Appendix A.4. Additionally, the *TIA1* p.A254G variant was also present in a single affected individual in the Project MinE database, while all other candidate mutations were absent from all additional ALS patient cohorts.

5.3.2 ALS-associated SNP variants

A total of 79 known population-based SNPs present in ExAC showed nominal statistical evidence of association with disease ($p<0.05$) when comparing allele counts between 81 FALS affected individuals (with an unidentified ALS mutation) and 60,706 ExAC control individuals. The results of Fisher's Exact testing are provided in Appendix A.3, Table A.3.3.1. After performing analysis to remove family bias as per

Table 5.2: **Novel non-synonymous variants identified in candidate genes among FALS and SALS patients.**

Gene	Cohort	Transcript accession	CHROM	POS	cDNA position	Amino acid change	Direct sequencing	Assessment of pathogenicity score*	Control database result	Present in additional ALS patient cohorts?
SPTBN4	FALS	NM_020971	19	41072150	c.6221G>C	p.R2074P	Validated	3.5574 - medium priority	Absent	no
EEF1D	FALS	NM_001960	8	144661974	c.834C>A	p.F278L	Validated	4.3154 - medium priority	Absent	no
ABCC2	FALS	NM_000392	10	101590549	c.2824G>A	p.D942N	Validated	0.5 - low priority	Absent	no
ABCC2	FALS	NM_000392	10	101595991	c.3558T>A	p.N1186K	Validated	2.8 - medium priority	Absent	
MTHFR	FALS	NM_005957	1	11852346	c.1621G>C	p.V541L	Validated	4.3108 - medium priority	Absent	no
DAGLB	FALS	NM_139179	7	6452495	c.1516G>A	p.E506K	NOT present	6.5378 - high priority	Present in one Latino individual	no
TIA1	FALS	NM_022037	2	70442597	c.761C>G	p.A254G	validated	7.1308 - high priority	Present in two SEA individuals	Project MinE (AC=1)
TIA1	SALS	NM_022037	2	70441601	c.881C>T	p.P294L	validated	5.0308 - high priority	Absent	no
TIA1	SALS	NM_022037	2	70457950	c.160C>A	p.H54N	validated	4.5308 - medium priority	Absent	no

Abbreviations: SEA; South East Asian; and AC. allele count.

*See Chapter 6, Section 6.2.3 for details of the pipeline for assessment of potential ALS pathogenicity.

Custom Script A.2.5, the association with disease remained nominally significant for 53 SNPs. Upon validation using Australian control cohorts, only 15 of these 53 SNPs showed nominally significant evidence of association with disease. Of these 15, seven were over-, and eight under-represented in FALS affected individuals (summarised in Table 5.3).

Re-analysis of association testing results from FALS using the Bonferroni corrected p-value of 6.5×10^{-5}, found that just seven SNP variants were significantly associated with FALS compared to ExAC controls (see Appendix A.3, Table A.3.3.1). Five of these seven SNPs were found to be the result of family bias (using Custom Script A.2.5). The two SNPs which withstood family bias testing (highlighted in Appendix A.3, Table A.3.3.1) were found to be Australian-associated variants after repeating testing with population-matched control cohorts. Therefore, no population-based SNPs were found to be significantly associated with FALS after Bonferroni correction.

Among 635 SALS affected individuals, 16 population-based SNPs within the four candidate genes (*CHCHD2*, *CHCHD3*, *CHCHD6* and *TIA1*) showed baseline statistical evidence of association with disease ($p<0.05$) when compared to 63,369 gnomAD controls of NFE descent (Appendix A.3, Table A.3.3.2). Of these associations, just two were replicated using Australian controls. An exonic *CHCHD3* variant was under-represented in SALS affected individuals while an intronic *TIA1* variant was over-represented in SALS affected individuals (Table 5.3). Using the Bonferroni corrected p-value of 5.20×10^{-6}, just two SNPs met significance compared to gnomAD NFE controls, however this significance was lost in each case when using Australian controls from MGRB. As such, no population-based SNPs were found to be significantly associated with SALS after Bonferroni correction.

Those SNPs in Table 5.3, shown to have baseline association with FALS or SALS ($p<0.05$), had Fisher's exact testing repeated using Project MiNE SALS affected individuals and control individuals. Results are presented in Appendix A.3, Table A.3.3.3. Just two variants, in *NEK1* (rs200161705) and *CNR2* (rs2501432), retained baseline significance ($p<0.05$) in this replication cohort. While the *NEK1* variant was over-represented in both affected individual cohorts, the *CNR2* variant was under-represented in our Australian FALS cohort, though over-represented in the Project MiNE SALS cohort.

Table 5.3: **Candidate gene SNPs potentially associated with ALS.**

Gene	CHROM	POS	rs ID	Cohort*	Patient MAF	ExAC/ /gnomAD MAF	DACC MAF#	MGRB MAF	Potential disease risk or protective allele?	Replicated in Project MiNE? ∧
SPTBN4	19	41060616	rs2242131	FALS	0.08	0.22	0.15	0.17	protective	no
SPTBN4	19	41071552	.	FALS	0.03	0.00	0.21	0.00	risk	no
C21orf2	21	45750145	rs11552066	FALS	0.02	0.16	0.11	0.12	protective	no
C21orf2	21	45759045	rs11870	FALS	0.09	0.34	0.50	0.22	protective	no
NEK1	4	170506525	rs200161705	FALS	0.02	0.00	0.00	0.00	risk	yes
EEF1A2	20	62124459	rs12480745	FALS	0.13	0.26	0.31	0.27	protective	no
EEF1A1	6	74227940	rs11556677	FALS	0.08	0.17	0.24	0.00	protective	no
BICD2	9	95483066	.	FALS	0.01	0.00	0.00	0.00	risk	no
BICD2	9	95526977	.	FALS	0.03	0.00	0.00	0.00	risk	no
CHCHD6	3	126676314	rs145020754	FALS	0.01	0.00	0.00	0.00	risk	no
CNR2	1	24201357	rs4649124	FALS	0.45	0.62	0.57	0.58	protective	no
CNR2	1	24201919	rs2502992	FALS	0.45	0.62	0.56	0.58	protective	no
CNR2	1	24201920	rs2501432	FALS	0.45	0.62	0.57	0.58	protective	yes
DAGLA	11	61507041	.	FALS	0.01	0.00	0.00	0.00	risk	no
KIF5A	12	57963020	rs181688415	FALS	0.04	0.01	0.02	0.01	risk	no
CHCHD3	7	132719349	rs78193687	SALS	0.06	0.07	N/A	0.08	protective	no
TIA1	2	70463334	rs78928004	SALS	0.00	0.00	N/A	0.02	risk	no

Refer to Appendix A.3, Tables A.3.3.1 and A.3.3.2 for p-value results comparing patient allele counts to each control cohort.
*FALS: 81 FALS patients from 69 families with an unidentified mutation; SALS: 635 SALS patients.
#N/A: intronic variant not covered in WES data available for DACC controls.
∧ Refer to Appendix A.3, Tables A.3.3.3 for p-value results.
Abbreviations: ExAC, Exome Aggregation Consortium; MGRB, Medical Genome Reference Bank.

5.4 Discussion

Among the 54 candidate ALS genes (Table 5.1) analysed in this Chapter, eight candidate mutations and 17 potentially disease-associated variants were identified. Further, an efficient pipeline was established for screening candidate genes in Australian FALS and SALS. As most of our FALS cohort included probands, and the SALS patients had no known relatives affected by ALS, there was little opportunity to analyse segregation of the novel non-synonymous candidate mutations. To support a pathogenic role of these candidate mutations, identification of additional unrelated patients with the identical, or other novel candidate mutations in the same gene, will be required. Given that the majority of known ALS gene mutations are found in less than 1% of patients, it is likely that a novel mutation would have a similar, or even lower frequency. As such, analysis of thousands of unrelated patients would be required to obtain adequate support for a causal role of one of these candidate ALS mutations, which is not possible in the Australian patient cohort alone. In an effort to gain such support, additional ALS patient cohorts including Project MiNE, the ALS database and the ALS variant server were accessed and screened for the candidate mutations. Unfortunately, this strategy only supported a causal role for one candidate mutation in *TIA1* (as discussed below).

The eight validated, novel non-synonymous variants identified in known ALS genes in FALS and SALS patients represent candidate ALS mutations. If any relatives of a proband who carries one of these candidate mutations, develops ALS in the future, they will be screened for the candidate mutation to establish if it segregates with disease (as described in Chapter 6). Further, newly recruited FALS probands and SALS patients will also be screened for these variants, or other candidate mutations in the same genes. This extends to a complete screen of all 54 candidate genes in the 635 SALS patient cohort. Unfortunately, as this cohort was ascertained in the latter stages of candidature, candidate gene screening was only possible for just four genes.

In silico assessment of potential pathogenicity suggested that three of the nine novel non-synonymous variants identified in NGS data, prior to validation, showed substantial functional similarity with known ALS genes. This suggested that these variants had the strongest potential to be causal ALS mutations. Two of these, which were both validated by Sanger sequencing, resided in the *TIA1* gene. This gene is a strong ALS candidate gene, as an RNA binding-protein with strong genetic support from FALS analysis and burden testing (Mackenzie et al., 2017). Further, functional

evidence supports a role for the mutated TIA1 protein in increasing the rate of phase transition, delaying stress granule disassembly and promoting accumulation of TDP-43 containing stress granules (Mackenzie et al., 2017). The *in silico* assessment of pathogenicity performed in this Chapter, showed that *TIA1* is highly expressed in both the brain and spinal cord, and has an average level of tolerance for genetic variation. Further, four novel non-synonymous *TIA1* variants have been identified in four individual Chinese SALS patients (Gu et al., 2018; Yuan et al., 2018; Zhang et al., 2018), adding further support to the causality of *TIA1* mutations in ALS. Together, this suggests that the candidate mutations in *TIA1* may cause ALS. However, only one of the candidate mutations, *TIA1* p.A254G, was consistently predicted to be damaging by protein prediction programs. This variant also showed evidence of conservation across species, suggesting a conserved evolutionary role. Notably, this same candidate mutation was identified in a FALS proband as well as a SALS patient from the Project MinE database. Therefore, of the two high priority candidate mutations identified in *TIA1*, p.A254G showed the strongest evidence for a causal role in ALS. It must be noted that two individuals of South East Asian descent form the gnomAD control database were also observed to carry this variant. However, this equates to a rare MAF of just 8.13×10^{-6} across all ancestries, and 6.50×10^{-5} in South East Asian individuals. It is possible that these individuals may go on to develop disease. As such, this candidate causal mutation showed mixed evidence for a role in the pathogenesis of ALS and further investigations are warranted.

The third candidate mutation that showed the most functional similarity to the known ALS gene mutations using the *in silico* pipeline was found in *DAGLB*. However, Sanger sequencing revealed that this variant was actually a false positive identification. Interestingly, the DNA sample used for WES of this FALS patient underwent WGA. Sanger sequencing of the WGA sample showed an inconclusive genotype, while that using the original, non-amplified sample showed a homozygous wild-type genotype (Figure 5.2). This highlights the pitfalls of WGA, as it has potential to introduce genetic variation to patient DNA samples. This stresses the importance of validating variants by Sanger sequencing, and caution when interpreting sequencing data derived from WGA DNA samples. Unfortunately, the second patient for whom WES data was generated from a WGA DNA sample did not have a non-amplified DNA sample available, and the candidate mutation identified in this patient (*EEF1D* p.F278L) remains to be fully validated. Chapter 8, Section 8.3.3 will provide a detailed discussion of artefacts in NGS data, including those arising from PCR amplification.

Both rare and common known population-based SNPs were tested for association with disease in FALS and SALS patients. The hypothesis was that known population-based SNPs in candidate ALS genes may confer disease-risk, or be in linkage disequilibrium with a nearby risk allele. In the case of FALS, risk alleles of strong effect could, in part, explain the reduced penetrance of disease within some families. In contrast, it is likely that analysis of SALS cohorts would reveal SNPs that confer a smaller disease-risk. Alternatively, population-based SNPs under-represented in patients may confer protection against the development of ALS.

In the current study, two approaches to significance were utilised when assessing association of population-based SNPs with ALS. The first was to simply use a standard significance threshold of $p<0.05$. Secondly, a Bonferroni correction was applied to account for multiple testing of all SNP variants identified by candidate gene screening in the given dataset, be it WES data from FALS with an unknown mutation (n genes=54; n total SNPs=741) or WGS data from SALS (n genes=4; n total SNPs=9,616). Application of the stricter, corrected significance threshold discarded any SNPs as being significantly associated with ALS in either cohort.

Given the nature of the analysis conducted here, a variety of different correction factors could have been chosen. As each candidate gene was screened and analysed as the evidence arose to implicate that gene in ALS, a correction factor accounting for the number of SNPs found within each gene in isolation may have been applied. As the number of SNPs identified in each gene varied greatly, such a corrected significance threshold would also have varied substantially for each gene when considered in isolation. Additionally, when screening sets of candidate genes, such as the seven Endocannabinoid system genes, this approach would arguably require correction for all SNPs identified across all of these genes. However, this would have resulted in even greater inconsistencies between the significance thresholds utilised for each of the 54 candidate genes analysed here. As such, Bonferroni correction for each gene was deemed too biased. Additionally, for those genes screened in both FALS and SALS, inconsistent significance thresholds would need to have been applied for the same gene in the two cohorts, as the FALS WES data identified far fewer variants annotated to each gene than SALS WGS data. Alternatively, correction could have considered only coding variants, or only those with control allele count data available. Again, these factors would have caused further inconsistencies between FALS WES and SALS WGS data, and between analyses of the same gene in the same ALS patient cohort but using different control cohorts, respectively. Given that Bonferroni correction removed all

significance, and analysis in this Chapter was completed on a gene-by-gene basis, with no *a priori* hypothesis, population-based SNPs with a p-value<0.05 were cautiously considered as potential disease-associated SNPs.

Association testing in FALS was complicated by the fact that the FALS cohort included some families with multiple affected individuals. As such, the inheritance of a SNP within a family may artificially inflate the allele frequency within this patient cohort. To overcome this limitation, an additional association testing strategy was incorporated into the FALS association testing regime, to determine whether an apparent association was an artefact of family bias. This was achieved by including a single member of each family, in turn, and repeating the statistical analysis. In order for an association to be considered significant, the p-value from each possible patient combination was required to reach the significance threshold. This strategy provided confidence that any statistically significant association was not merely an artefact of the family bias effect. Indeed, we found that 34.6% of nominally significantly FALS-associated SNPs were attributable to such family bias, highlighting the importance of accounting for this confounding variable in genetic association analyses. Unrecognised familial relationships within a patient cohort are likely to introduce a level of bias to any genetic association study. It is possible that many reported ALS associated variants are artefacts of distant family relationships, which may underlie the failure to replicate some genetic association studies in ALS. Unfortunately, by including only a single member of each family, the statistical power of the association analysis was reduced. Ideally, to retain maximal statistical power while avoiding the introduction of family bias, this analysis would have accounted for relatedness between individuals within the cohort. To do so, a baseline degree of relatedness between a comparably sized cohort of ancestrally matched, unrelated individuals would need to be established. However, such an analysis would require access to individual level genotype data from such a control cohort, which was not available within the time constraints of this candidature.

In addition to accounting for family bias, this study also sought to account for bias due to ancestry. As discussed in Chapter 4, the Australian population has a diverse ethnic background, and therefore is likely to possess a unique genetic architecture. Association analysis was therefore repeated using two separate Australian control cohorts (DACC and MGRB). Both cohorts are relatively small (consisting of 967 and 1,144 individuals respectively) when compared with the international control databases ExAC and gnomAD (each consisting of tens of thousands of individuals).

For this reason, the discovery phase of association analysis was conducted using the larger, international cohorts to increase statistical power, while the Australian cohorts were used for validation purposes. Our analyses showed that 71.7% of (non-family biased) FALS associated SNPs, and 87.5% of SALS associated SNPs, were not replicated using Australian control cohorts. This suggests that these associations may be reflective of ancestry rather than disease state. This highlights the paramount importance of ensuring patient and control cohorts are derived from the same population. Accordingly, it is important to note that the gnomAD controls of NFE descent were used in the SALS association analyses, as this was the most abundant ancestral background among the Australian ALS patient cohort. Yet still, this control cohort showed numerous allele frequency differences to Australian-based control cohorts. The case may be that the failure to replicate these variant associations was simply due to random sample variation causing allele frequency differences between the control cohorts. A potential method to determine whether allele frequencies were influenced by Australian ancestry, would be to compare the control datasets in a pseudo case-control association analysis, to determine whether the Australian control cohorts could be distinguished from the mixed ancestry and European cohorts, as well as from each other. It must also be noted that the use of multiple control cohorts in itself introduces an increased burden of multiple testing. As such, the significance thresholds utilised here may be too lenient, therefore these results must be treated with caution.

The statistical power of the association analyses presented in this chapter are quite small. This a result of the availability of small sample sizes of the case cohorts, being just 81, 61 and 635 for FALS, FALS proband and SALS patients, respectively. The sample size of the control cohort also contributes to the degree of statistical power, thus the association analyses using the ExAC or gnomAD control databases had superior power to those analyses using the smaller Australian control cohorts from MGRB and DACC. In addition to sample size, statistical power is further influenced by a variety of factors including disease prevalence, risk-allele frequency, linkage disequilibrium and the inheritance model the risk allele(s) (ie. additive, dominant, multiplicative) and risk-allele effect size (Hong and Park, 2012). Given that allele frequency and linkage disequilibrium values are unique to each SNP, and the uncertainty of the inheritance model and effect size of ALS risk-alleles (and the potential of these to also be unique to each risk-allele), it was not possible to perform formal power calculations. As such, rather than actually identifying real ALS risk-alleles, our approach acts as a tool for identifying potentially interesting risk-alleles that warrant detailed analysis

for association with ALS in larger patient cohorts, under various assumptions around inheritance models and effect sizes.

Seventeen SNPs were considered as being nominally significantly associated with disease in the FALS and SALS cohorts ($p<0.05$), and implicate the presence of risk alleles in 12 different genes. The Australian sample cohorts (81 FALS and 635 SALS) were relatively small, though the unique genetic background of this population provides a rare resource for investigating ALS-associated variants. Replication of these results using the Project MiNE case-control cohort supported the over-representation of the *NEK1* rs200161705 SNP in ALS patients compared with controls, suggesting this variant may be an interesting ALS risk allele. Interestingly, the *CNR2* rs2501432 was under-represented in Australian FALS compared with control individuals from ExAC, DACC and MGRB, however in the Project MiNE case-control cohort, this variant showed significant over-representation in SALS patients compared with controls. While this result may simply represent a false positive finding in either one or both of the case cohorts, it is possible that it may reflect a difference between the genetic architecture of FALS and SALS. Alternatively, it may indicate that some sort of complex population stratification effect is at play, in that this variant does contribute to ALS phenotypes, but its effect is dependent on interactions with other genetic variants. That is, on an Australian-based genetic background, this variant may confer protection against ALS, while on a European-based genetic background this variant may increase the risk for ALS. Together, these association results warrant further investigations of these variants in larger cohorts with more diverse genetic backgrounds.

"You can't just give up, is that what a dinosaur would do?"

Joey Tribbiani - Friends

6

Novel disease gene discovery in ALS families

6.1 Introduction

This Chapter addresses the first part of Aim 3 of this project; to identify novel ALS genes and mutations in families with a history of ALS. The analyses presented here centre around five ALS families (FALSmq28, FALS15, FALS45, FALSmq2 and FALSmq20), each exhibiting reduced disease penetrance. DNA samples were available from just two or three informative members from each family, meaning that these families have limited genetic power. These small families are not amenable to traditional linkage analysis methods alone, in contrast to those larger ALS families that were utilised by our laboratory to discover the known ALS genes *TARDBP* (Sreedharan et al., 2008), *UBQLN2* (Deng et al., 2011) and *CCNF* (Williams et al., 2016b). A combination of next-generation sequencing (NGS), bioinformatic analysis, and genome-wide linkage analysis was employed to identify a list of candidate mutations in each of the five families. A multi-faceted *in silico* pipeline was developed to functionally characterise each candidate mutation in order to assess its potential pathogenicity, relevance to ALS, and ultimately prioritise those warranting downstream *in vitro* analysis.

6.1.1 ALS gene discovery and disease aetiology

Genetic discoveries in ALS over the past 25 years have laid the foundation for the majority of our current understanding of the disease biology underlying ALS. As established in Chapter 1, Section 1.4, at least 25 genes have been found to harbour ALS causal mutations. These genes, both in isolation and collectively, have served to highlight the role of specific molecular pathways and mechanisms contributing to ALS pathogenesis. Thus, ALS gene discoveries have also provided the targets for downstream research into the pathogenic underpinnings of disease.

The critical links between the wider biology of ALS and the genetic causes of disease have been evident since the first discovery of causal mutations in *SOD1* (Rosen, 1993). Soon after this genetic discovery, the SOD1 protein was identified within protein aggregates found in affected motor neurons in patient tissue (Shibata et al., 1996). The discovery of *SOD1* mutations in ALS also lead to the first indication of the role oxidative stress plays in ALS pathogenesis. Later, following the identification of the TDP-43 protein as a major constituent of hallmark ubiquitinated neuronal cytoplasmic inclusions (Arai et al., 2006; Neumann et al., 2006), causal mutations in the gene encoding this protein, *TARDBP*, were identified in familial and sporadic ALS cases (Sreedharan et al., 2008). This again demonstrated a critical link between ALS genetics and understanding the pathogenesis of disease, and was among the first clues to implicate the role of RNA-binding proteins and RNA-processing pathways in ALS. Other ALS genes including *FUS* and *UBQLN2* have continued to inform our understanding of ALS pathogenesis, having also been found within the hallmark protein inclusions (Deng et al., 2011; Neumann et al., 2011). As the number of ALS genes has grown, numerous genes have clustered together to implicate common molecular pathways in ALS, most notably RNA-processing and protein homeostasis (discussed in Chapter 1, Section 1.3.5). Further, most animal and cell models of ALS have been developed through the introduction of known ALS gene mutations.

Known ALS gene mutations account for approximately 60% of Australian FALS, leaving almost 40% of these FALS cases to be solved (Paper I, McCann et al., 2017). Additionally, just 5% of SALS patients harbour a known ALS mutation (Renton et al., 2014). Therefore, the cause of disease in over 90% of ALS patients remains unknown, meaning that numerous genetic mutations/perturbations that cause or contribute to ALS are yet to be discovered. Each novel ALS gene discovery will provide a new opportunity to further our knowledge and understanding of the biology underlying disease, as has been achieved for those ALS genes already identified. The identification

of novel ALS genes from family studies will benefit both FALS and SALS patients alike, through an increased understanding of disease aetiology.

6.1.2 Approach to novel disease gene discovery

As established in Chapter 1, Section 1.5.2, the wide-spread adoption of NGS technologies, particularly whole exome- (WES) and whole-genome (WGS) sequencing, has facilitated substantial growth in the number of identified disease genes for ALS and many other hereditary conditions. The majority of this success can be attributed to the application of family-based filtering strategies to NGS datasets, namely segregation analysis and filtering of common population-based variants. The effectiveness of this approach has been further enhanced when coupled with traditional genetic linkage analysis techniques in large families. Indeed, our laboratory successfully applied this strategy to identify novel ALS genes including *CCNF* (Additional Paper I; Williams et al., 2016b) and *UBQLN2* (Williams et al., 2012a). These studies used large ALS families with multiple affected individuals over several generations, together with unaffected "married-in" spouses and/or parents, providing sufficient genetic power for effective segregation and genetic linkage analysis. However, limited sample availability from the unsolved ALS families limits the application of these analyses (discussed in Chapter 1, Section 1.6.1). In cases where segregation analysis is applied to just two first-degree relatives (i.e. relatives whom share an average of 50% of their DNA sequence), a long list of shared variants (ie. variants identical by descent), or candidate mutations, is expected. In the absence of genetic data from additional informative family members, alternative strategies are required to elucidate the causal mutation in such small ALS families.

A number of commonalities are evident among the known causal ALS gene mutations, on both a gene and variant level. These include a range of biological and genetic characteristics (discussed in Chapter 2, Section 2.5). Many of these can be assessed using *in silico* tools or databases (see Table 2.5). Therefore, as part of this Chapter, an *in silico* strategy was developed to assess the potential pathogenicity of a given candidate mutation in an ALS family, utilising these characteristics together with their associated *in silico* tools. This approach was used to prioritise candidate mutations with the highest potential pathogenicity in each family.

6.2 Methods

6.2.1 ALS families

All available family members from families FALSmq28, FALS15, FALS45, FALSmq2 and FALSmq20 were recruited and had their DNA samples collected according to Chapter 2, Section 2.1.1. Each individual also provided informed written consent according to Chapter 2, Section 2.1.2. Pedigrees are provided in Figures 6.1 – 6.5.

Each family member was classified as either an ALS affected individual/patient, obligate mutation carrier, "married-in" control or "at-risk" individual as described in Chapter 2, Section 2.1.3.1. Among these subject types, affected individual/patient, obligate mutation carrier and "married-in" control parent of an affected individual were all considered informative for family-based genetic analysis.

DNA was available from multiple informative members from the five ALS families studied here. Each family was negative for known ALS gene mutations. Individuals from families FALS15, FALS45, FALSmq2 and FALSmq20 underwent known ALS gene screening as part of Paper I (Chapter 4, Section 4.3.1). FALSmq28 patients were recruited after publication of this article, however they too underwent known ALS gene screening following identical protocols.

Table 6.1 describes the DNA sequencing data available for each family to facilitate novel gene discovery. Detailed methodology for each technology type are provided in Chapter 2.

TABLE 6.1: **Summary of available data from multi-generation families.**

	Number of DNA samples available				**Available datasets**		
	Affected	**Obligate mutation carriers**	**"Married-in" controls#**	**"At-risk" individuals**	**WES**	**WGS**	**SNP microarray genotypes**
FALSmq28*	2	1	2	11	yes	yes	yes
FALS15	1	1	0	0	yes	.	.
FALS45	2	0	0	0	yes	.	.
FALSmq2	1	1	0	0	yes	.	.
FALSmq20	2	0	0	0	yes	.	.

*Only the three mutation carriers (i.e. affected and obligate mutation carriers) from FALSmq28 underwent WES and WGS, while all family members underwent SNP genotyping.

All "married-in" control individuals were the unaffected parent of an ALS patient.

FALSmq28

FALSmq28 is a six-generation family, in which two male individuals were both diagnosed with ALS. These two affected individuals were second cousins. DNA from both affected individuals, and the mother of one affected individual (an obligate mutation carrier), each underwent WES and WGS. These three informative individuals, and an additional 11 "at-risk" family members and two "married-in" controls also underwent SNP genotyping for genome-wide linkage analysis. The pedigree for this family is provided in Figure 6.1.

Clinical information

Both affected individuals had classical ALS with no evidence of FTD. The proband had bulbar onset at 51 years of age, having presented with slurred speech. He was formally diagnosed with ALS 11 months later at 52 years of age. After a disease duration of 58 months, the proband passed away at 56 years of age. The second affected individual also displayed his first symptoms at 51 years, though with spinal onset and symptoms in his left leg. He too was formally diagnosed at 52 years, and is currently alive at 55 years of age, having lived with disease for 48 months.

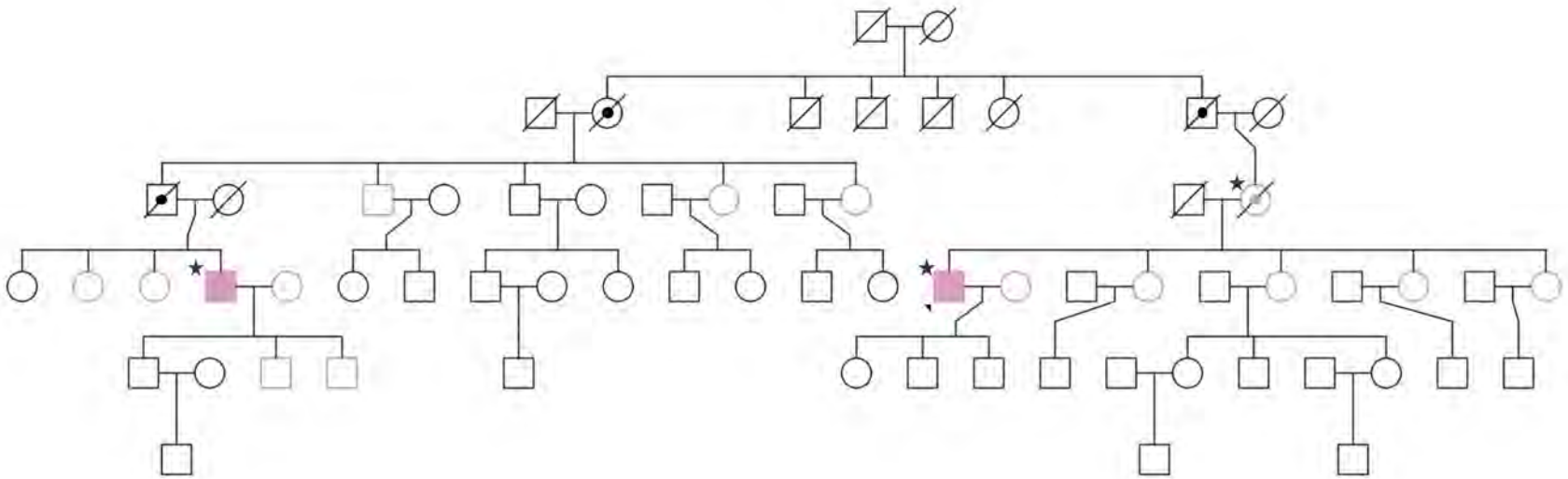


FIGURE 6.1: **Pedigree of family FALSmq28.** FALSmq28 is a six-generation family with a history of classic ALS. Females are indicated by circles, males by squares. Filled symbols indicate individuals affected by ALS. Symbols containing a small filled circle indicate obligate mutation carriers, unaffected by ALS. A diagonal strikethrough indicates a deceased individual. Pink coloured symbols indicate individuals with DNA samples available, and all of these individuals underwent SNP microarray genotyping. Individuals marked with purple stars underwent WES and WGS. The black arrow represents the family proband.

FALS15

FALS15 is a four-generation family from which a male proband and his first cousin-once-removed were both diagnosed with ALS. WES was performed for the proband and his mother, an obligate mutation carrier. The pedigree for this family is provided in Figure 6.2.

Clinical information

Limited clinical details were available from this family. Both affected individuals were diagnosed with classical ALS with no evidence of FTD. The proband presented with bulbar symptoms at 58 years of age.

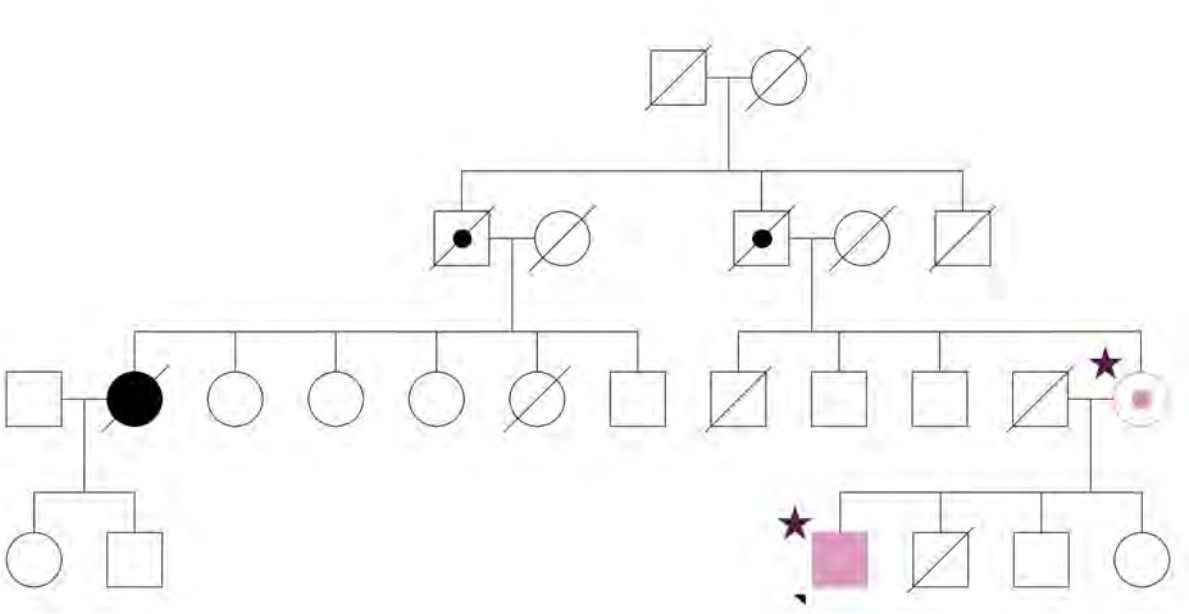


Figure 6.2: **Pedigree of family FALS15.** FALS15 is a four-generation family with a history of classic ALS. Females are indicated by circles, males by squares. Filled symbols indicate individuals affected by ALS. Symbols containing a small filled circle indicate obligate mutation carriers, unaffected by ALS. A diagonal strikethrough indicates a deceased individual. Pink coloured symbols indicate individuals with DNA samples available. Individuals marked with purple stars underwent WES. The black arrow represents the family proband.

FALS45

FALS45 is a four-generation family consisting of a male proband whose father was also affected by ALS. WES was carried out for the proband and his mother, who was a "married-in" control. The pedigree for this family is provided in Figure 6.3.

Clinical information

Limited clinical details were available from this family. Both affected individuals were diagnosed with classical ALS with no evidence of FTD.

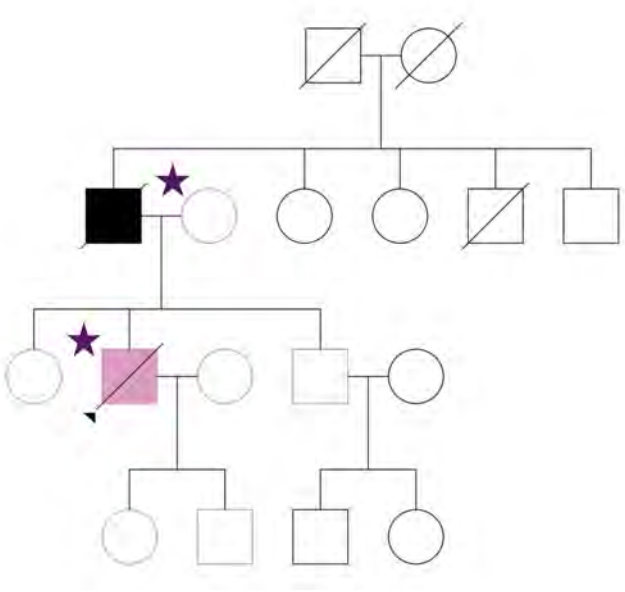


FIGURE 6.3: **Pedigree of family FALS45.** FALS45 is a four-generation family with a history of classic ALS. Females are indicated by circles, males by squares. Filled symbols indicate individuals affected by ALS. A diagonal strikethrough indicates a deceased individual. Pink coloured symbols indicate individuals with DNA samples available. Individuals marked with purple stars underwent WES. The black arrow represents the family proband.

FALSmq2

FALSmq2 is a four-generation family. The male proband affected individual had a maternal uncle who was also diagnosed with ALS. WES was performed for the proband and his mother, who was an obligate mutation carrier. The pedigree for this family is provided in Figure 6.4.

Clinical information

A history of classical ALS with no evidence of FTD was reported for this family. The proband presented with an affected right arm at 50 years of age, and was formally diagnosed with ALS six months later. This affected individual died at 54 years of age, 51 months after diagnosis.

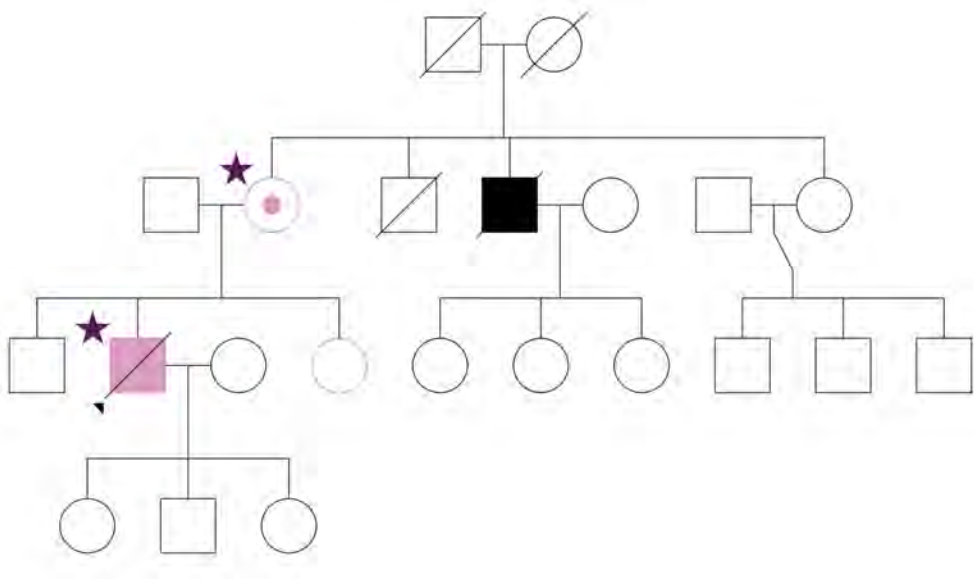


Figure 6.4: **Pedigree of family FALSmq2.** FALSmq2 is a four-generation family with a history of classic ALS. Females are indicated by circles, males by squares. Symbols containing a small filled circle indicate obligate mutation carriers, unaffected by ALS. A diagonal strikethrough indicates a deceased individual. Pink coloured symbols indicate individuals with DNA samples available. Individuals marked with purple stars underwent WES. The black arrow represents the family proband.

FALSmq20

FALSmq20 is a four-generation family. A male proband patient and his mother were both diagnosed as ALS patients. WES was performed for both affected family members. The pedigree for this family is provided in Figure 6.5.

Clinical information

The proband presented with symptoms in the right arm at 40 years of age, while his mother experienced bulbar onset at 75 years of age. Both received formal diagnoses after approximately two years. The affected mother of the proband also displayed symptoms of dementia. The proband was alive at the time of analysis and has lived with ALS for 143 months, however his mother died at 78 years of age after a disease course of just under 46 months.

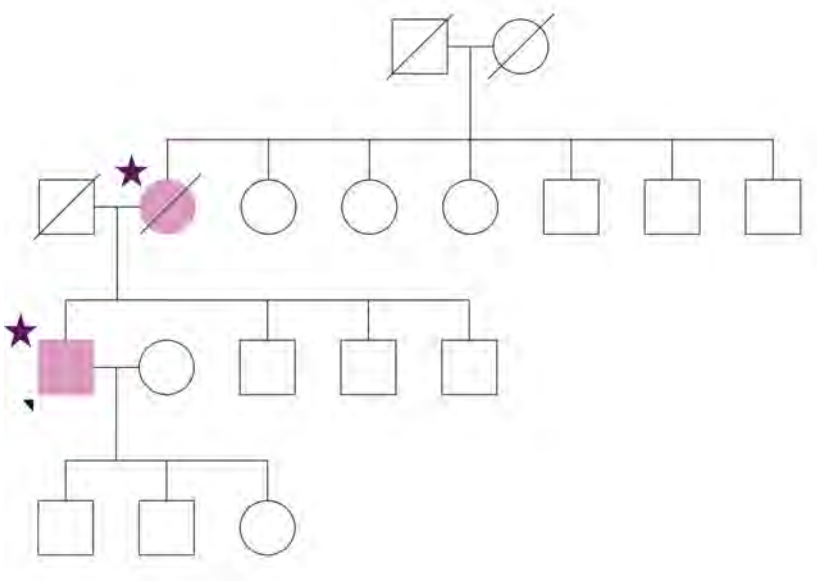


FIGURE 6.5: **Pedigree of family FALSmq20.** FALSmq20 is a four-generation family with a history of classic ALS. Females are indicated by circles, males by squares. Filled symbols indicate individuals affected by ALS. A diagonal strikethrough indicates a deceased individual. Pink coloured symbols indicate individuals with DNA samples available. Individuals marked with purple stars underwent WES. The black arrow represents the family proband.

6.2.2 Identifying candidate ALS causal mutations in each family

In order to identify the ALS causal gene mutation in each family, a custom family analysis approach was employed. The families described above represent two different family types. Firstly, a medium-sized family with three informative, and 13 additional DNA samples available (FALSmq28), and secondly, smaller families with just two informative DNA samples available from a parent-offspring pair (FALS15, FALS45, FALSmq2 and FALSmq20). The basic shared variant and filtering analysis pipeline applied to both family types is described in Table 6.2. Briefly, NGS data from each family was filtered to identify genetic variants shared by all affected and obligate mutation carrying family members, and absent from any "married-in" control individuals (ie. variants identical by descent). These variants were then visually inspected using NGS read data to confirm the nucleotide identity of the SNP call. Any variants found to have an incorrect nucleotide identity were corrected, and also had their control

Table 6.2: **Basic steps of family-based analysis pipeline for gene discovery.**

Analysis Step	**Description**	**Relevant Section or Custom Script**		
		Small families - WES	**FALSmq28 - WES**	**FALSmq28 - WGS**
Original NGS VCF	VCF(s) containing WES or WGS data for individual family members	Section 2.2	Section 2.2	Section 2.2
Generate family VCF	VCF containing all family members, and only called variants with an alternate allele present in at least one affected family member	Script A.2.10	Script A.2.12	Section 3.5.2
ANNOVAR annotation	Append bioloogical annotations for each variant	Section 2.2.2.2	Section 2.2.2.2	Section 2.2.2.2
Custom family analysis	Retain all variants shared by affected individuals (and obligate carriers) and remove any variants present in unaffected "married-in" control individuals	Script A.2.11	Script A.2.13	Script A.2.14
First-tier filtering	Remove common variants in the general population, using dbSNP, 1000Genomes, ExAC and/or gnomAD NFE MAF>0.0001			
	Remove non-coding variants (using Func.refGene annotations)			
	Remove synonymous variants (using ExonicFunc.refGene annotations)			
First-tier validation	Visualise WES reads using IGV to confirm genotype calls	Section 2.4.1	Section 2.4.1	Section 2.4.1
	Ammend ExAC/gnomAD MAF and AC for variants with incorrect allele calls	N/A	N/A	N/A
Second-tier filtering	Remove less common variants in the general population (i.e. variants with AC>2 across ExAC and gnomAD)	Section 3.5.3 & Script 3.11	Section 3.5.3 & Script 3.11	Section 3.5.3 & Script 3.12
	Remove less common variants in the Asutralian population (i.e. variants with AC>2 across ExAC, gnomAD, DACC and MGRB)			
Second-tier validation	Confirm variant is present in each affected family member, and/or absent from any "married-in" controls	Section 2.4.2	Section 2.4.2	Section 2.4.2

Abbreviations: VCF, variant call file; ExAC, Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MAF, minor allele frequency; IGV, Integrative Genomics Viewer; AC, allele count; and MGRB, Medical Genome Reference Bank.

database values (ie. minor allele frequency: MAF and alternate allele count: AC) corrected. All variants were then filtered to remove any common, rare, or Australian population-based variants. Any remaining variants were then validated using Sanger sequencing to confirm their presence in affected individual DNA samples, and absence from "married-in" control DNA samples. The following sections provide specific details of these analysis steps, which varied slightly between the two different family types due to the availability of family samples (and consequently sequencing data), and control database versions at the time of analysis.

6.2.2.1 FALSmq28

As described above, WES, WGS and SNP microarray genotype data were available for FALSmq28 (Table 6.1). Pedigree analysis showed male-to-male transmission, therefore an autosomal inheritance model was assumed. However, the dominant inheritance of disease was inconclusive, thus both dominant and recessive inheritance (and therefore both heterozygous and homozygous variants) were considered. However, where a conclusive disease model was required, an autosomal dominant inheritance model was assumed.

A combination of NGS shared variant analysis and SNP microarray genome-wide linkage analysis was employed for novel disease gene discovery in FALSmq28. NGS data processing and filtering was achieved using Custom Scripts developed as part of this thesis as described in Table 6.2 (with each being provided in the Appendix). Figure 6.6 outlines the analysis pipeline for this family, which consisted of three complementary analysis strategies as described below, each of which was applied separately to the WES and WGS datasets generated for this family. The initial analysis phase (Analysis 1; 6.6A) considered only coding variants. Genome-wide linkage analysis (as described below) then followed. The results of genetic linkage analysis were subsequently used to refine the genomic regions included in NGS shared variant analysis. Analysis phases 2 and 3 were restricted to genomic regions with logarithm of odds (LOD) scores >0 and >-2, respectively. The Custom Script 3.7 was used to extract these regions from the complete family VCF prior to standard filtering, similar to Analysis 1. Both Analyses 2 and 3 considered variants falling within all functional classes of the genome other than intronic and intergenic regions. Analysis 3 also employed three additional filtering steps to reduce the variants under analysis to a manageable number. Quality filtering was applied prior to basic family-based filtering to remove any variants with a genotype quality (GQ) score <20 in WES or WGS data from all three informative

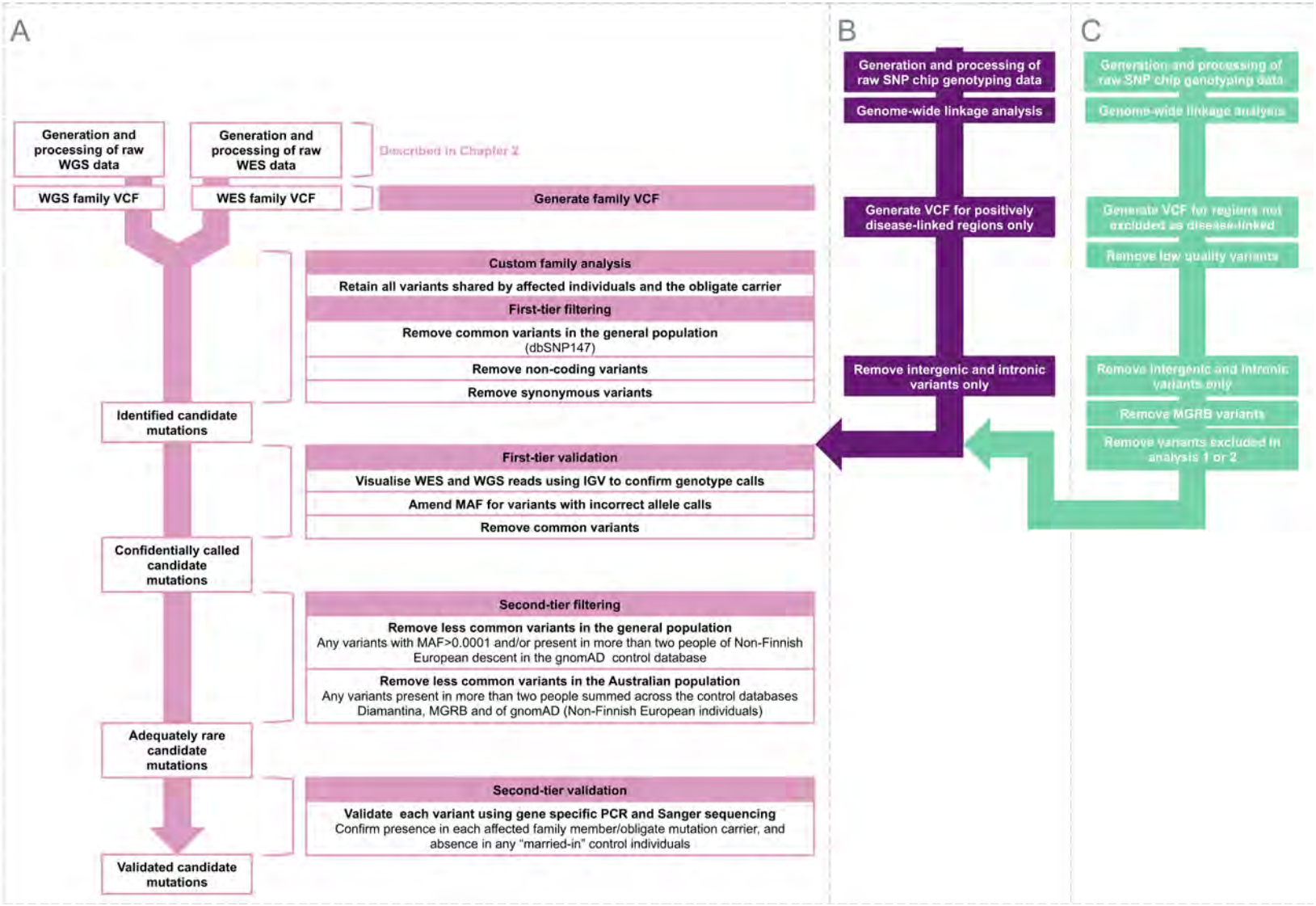


Figure 6.6: **Novel gene discovery analysis pipeline for family FALSmq28.** Overview of the analysis, filtering and validation steps applied to NGS data from the medium-sized family FALSmq28 to identify the causal ALS mutation in this family. A) Analysis 1. A traditional shared variant analysis approach was employed to identify coding candidate mutations. Additional steps and amendments were then applied to Analysis 1, to perform extended shared variant analysis. B) Analysis 2. This analysis included genomic regions showing the strongest evidence of linkage to disease (i.e. genomic regions with LOD scores >0 from genome-wide linkage analysis) and included all coding or regulatory variants falling within these regions. C) Analysis 3. This analysis included any genomic regions possibly linked to disease (i.e. genomic regions with LOD scores >-2 from genome-wide linkage analysis), high quality variants (those with a GQ>20 in all family members) and included all coding or regulatory variants falling within these regions.

family members (using the Custom Script 3.1). Secondly, Australian variants present in the MGRB database were removed (as described for Australian variant filtering in Table 6.2) prior to first-tier validation, in order to reduce the variants requiring visual inspection. Thirdly, any variants which had already been excluded by Analysis 1 or 2 were also removed before further filtering steps were applied.

Genome-wide linkage analysis

Raw data from SNP microarrays were processed according to Chapter 2, Section 2.3 by the service provider, Macrogen (Korea). The service provider also completed consistency and integrity checks for the output files using PedStats (Wigginton and Abecasis, 2005), and SNP pruning for those variants with a MAF<0.001 or in strong linkage disequilibrium. The data received from the service provider included a ped file containing family information and genotype data (example in Appendix A.4, Figure A.2), a dat file describing the structure of the ped file (example in Appendix A.4, Figure A.3) and a map file defining the chromosomal location of each SNP marker (example in Appendix A.4, Figure A.4).

The ped file was first amended to assign the correct affection status and append liability classes (described in Table 6.3) for each family member, using the Custom R Script in Appendix A.2.15. The ped, dat and map files were then separated out for each chromosome, to facilitate linkage analysis by chromosome, using the Custom R Script in Appendix A.2.16.

Genome-wide parametric linkage analysis was performed using Merlin software (Version 1.1; Abecasis et al., 2002). This assumed an autosomal dominant disease model with a disease allele frequency of 0.0001, and liability classes for age-dependent penetrance as shown in Table 6.3 (specified using the model file shown in Appendix A.4, Figure A.5). Linkage analysis was then completed separately for each chromosome using the Custom Script in Appendix A.2.17. This script utilised options to specify equal allele frequencies, reduce memory requirements and output results in a tabular format. Analysis and plotting of LOD score results was completed using R and the ggplot2 package (using the Custom Script in Appendix A.2.18).

Table 6.3: **Liability classes for linkage analysis, based on the age-dependent penetrance of ALS.**

Age group (years)	<30	31 - 40	41 - 50	51 - 60	61 - 70	70+
Heterozygote penetrance	0.01	0.3	0.5	0.7	0.85	0.9
Homozygote penetrance	0.01	0.3	0.5	0.7	0.85	0.9

Distribution of variants across the genomic functional classes

In order to determine how the variants identified by WES and WGS were distributed across the different functional classes of the genome, the Custom Script in Appendix A.2.19 was developed. This was applied to both general functional classes (downstream, exonic, exonic;splicing, intergenic, intronic, ncRNA_exonic, ncRNA_exonic;splicing, ncRNA_intronic, ncRNA_intronic;splicing, splicing, upstream, upstream;downstream, UTR3; UTR5) and exonic functional classes (frameshift deletion, frameshift insertion, nonframeshift deletion, nonframeshift insertion, nonsynonymous SNV, stopgain, stoploss, synonymous SNV, unknown), for both WES and WGS sequencing datasets.

6.2.2.2 Small families

As described above, WES data was available for two informative family members from each of the families FALS15, FALS45, FALSmq2 and FALSmq20. Though limited clinical data was available for the ancestors of each proband patient, pedigree analysis showed no evidence of consanguineous unions or a lack of male-to-male transmission in any family. Additionally, relatedness analysis using KING (v2.1; Manichaikul et al., 2010) has previously shown the relationship coefficients between relatives are as expected, meaning that no underlying consanguinity exists in these families (unpublished data). As such, an autosomal dominant model of inheritance was assumed for each family. The analysis pipeline applied to each of these four families is provided in Figure 6.7. The Custom Script in Appendix A.2.11 was used for bioinformatic filtering.

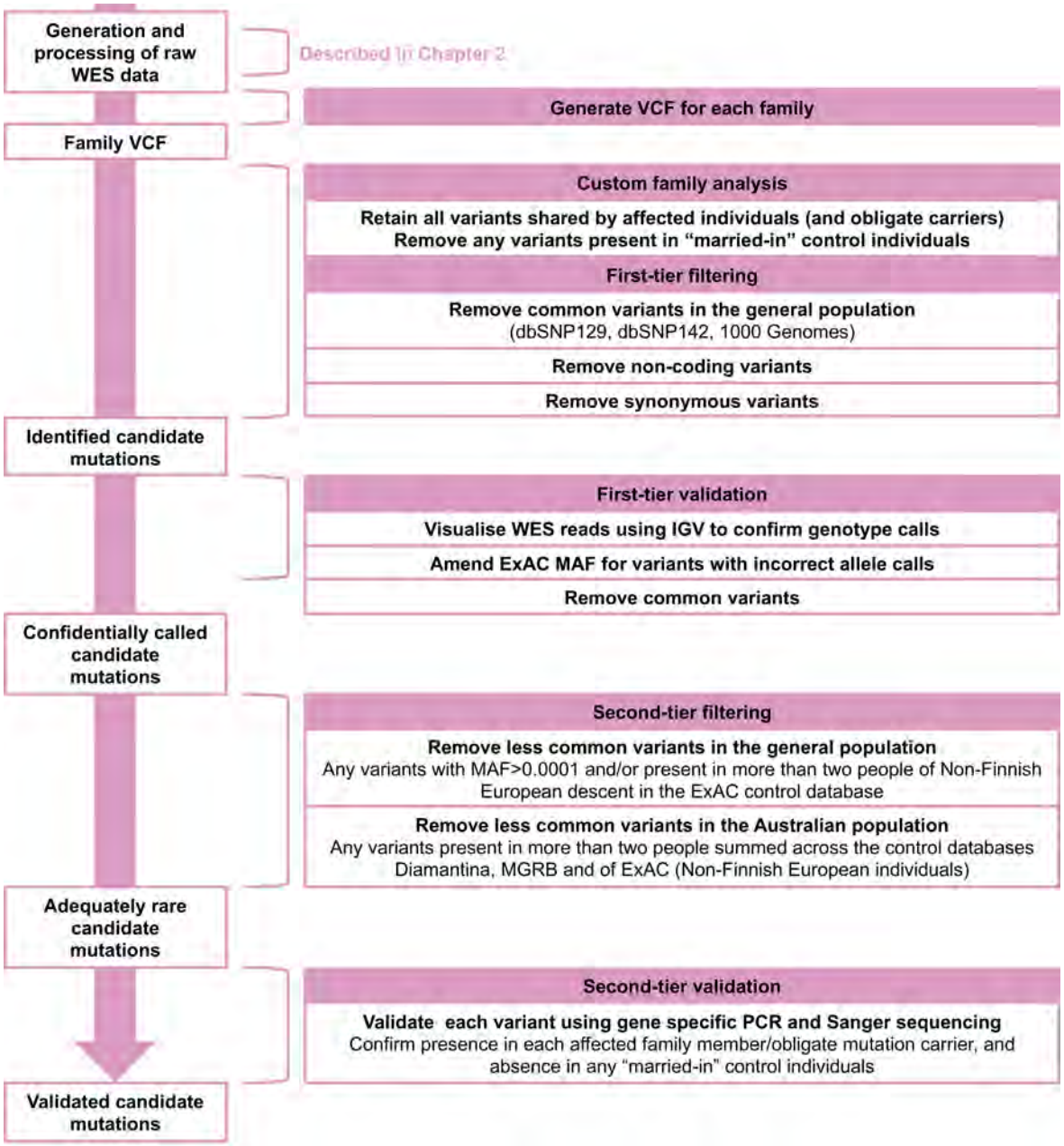


FIGURE 6.7: **Small family novel gene discovery analysis pipeline.** Overview of the analysis, filtering and validation steps applied to families FALS15, FALS45, FALSmq2, and FALSmq20 for novel disease gene discovery.

6.2.3 Assessment of potential for ALS pathogenicity using *in silico* tools

Following family analysis and filtering, each candidate mutation was functionally characterised and assessed for potential ALS pathogenicity using a combination of *in silico* tools. This included evaluation of gene specific characteristics such as expression levels and tolerance for genetic variation, as well as variant specific characteristics relating to predicted functional consequences and conservation across species (details provided in Chapter 2, Section 2.5, Table 2.5).

Based on the results of these *in silico* assessments, a scoring system was developed to rank all candidate mutations from each family according to their apparent potential for ALS pathogenicity. Table 6.4 describes this scoring system. In order to validate the scoring system, it was applied to the 11 causal ALS point mutations present in our FALS patient cohort (as established in Paper I, Chapter 4, Section 4.3.1) and three common population-based SNP variants.

Genic xtolerance formula explained

The formula devised to score genic tolerance was based on two metrics, RVIS percentile scores and ExAC missense z-scores. Therefore, these two metrics were each assessed as a separate component of the genic tolerance formula, with these two components then being added together to obtain an overall genic tolerance score. Genic tolerance has consistently been reported as a better predictor of pathogenicity compared with gene expression, protein predictions and amino acid conservation across species (MacArthur et al., 2014; Petrovski et al., 2013; Richards et al., 2015). As such, as part of the *in silico* scoring system developed here, genic tolerance was weighted more highly than the other characteristics. Specifically, genic tolerance contributed to four points of the overall *in silcio* assessment of pathogenicity score while all other characteristics each contributed two points of the overall score. The RVIS percentile scores and ExAC missense z-scores contributed equally to the genic tolerance score, and thus each contributed two points to this score.

RVIS scores assess whether a gene has more or less common functional variation relative to what is expected, given its level of neutral variation (Petrovski et al., 2013). Across the genome, these scores have a normal distribution. The RVIS percentile score (`http://genic-intolerance.org/`; Petrovski et al., 2013) was chosen as it was relative across all known human genes, was always positive (ranging from zero

to 100), and therefore would not produce a negative score. As this represented the percentile of most intolerant genes the gene of interest fell within (i.e. the smaller the percentile value, the more intolerant the gene) the inverse value was used to score more intolerant genes more highly. The inverse value was then multiplied by two so that the RVIS metric contributed two of the four genic tolerance points.

The ExAC missense z-scores were calculated by comparing the number of expected missense variants (based on the size of the gene) to the number of observed missense variants for each gene in the complete ExAC control dataset and represents the number of standard deviations the observed value was from the expected value (http://exac.broadinstitute.org/; Lek et al., 2016). Negative z-scores indicated more variants than expected (increased tolerance), and positive z-scores indicated less variants than expected (decreased tolerance). This model assumed a normal distribution, thus 99.994% of all observed values will fall within four standard deviations of the expected value. Therefore, the z-scores were divided by four to obtain a relative metric. Again, this value was multiplied by two so that the ExAC missense constraint score contributed to two of the four genic tolerance points. However, ExAC missense scores across the genome range from -8.64 to 13.88, therefore negative scores and scores greater than 2 were possible. To account for this in the scoring system presented here, score thresholding was applied. That is, any negative values were corrected to zero, so that genic intolerance did not mask scores from the other characteristics in the overall pathogenicity assessment score, and any scores greater than two were rounded down to two, to avoid inflation of the overall pathogenicity assessment score purely based on genic intolerance.

6.2.4 Additional evidence supporting potential pathogenicity

In silico analyses

Additional gene/variant characteristics were analysed as complementary traits to further support or refute the potential pathogenicity of each candidate mutation. This included gene/protein descriptions, known links to neurodegenerative disease, protein structure, protein interacting partners and the addition/removal of post-translational phosphorylation sites (details provided in Chapter 2, Section 2.5, Table 2.5).

Additional ALS patient cohort screening

Various other national and international ALS patient cohorts were also examined for the

presence of each candidate mutation. All such cohorts consisted of individuals of European ancestry, in accordance with the ancestry of the Australian ALS families under analysis. These cohorts included our in-house WES dataset from FALS affected individuals with an unidentified ALS mutation (n=81; 61 probands), and WGS dataset from 635 SALS affected individuals (within the 850-sample VCF). Further, WES/WGS data was obtained from dbGAP (https://www.ncbi.nlm.nih.gov/gap/) for 247 FALS affected individuals (dbGaP Study Accession: phs000101.v5.p1). These three datasets were screened using Custom R Scripts (Appendices A.2.4, A.2.6 and a variation of line 54 of Appendix A.2.4, respectively). Three publicly available datasets were also screened using their web browser interfaces. These were Project MiNE (n=4,366 SALS WGS; http://databrowser.projectmine.com/), the ALS data browser (ALSdb; n=2,800 FALS and SALS WES; http://alsdb.org/) and the ALS variant server (AVS; n=1,138 FALS and 277 SALS WES; http://als.umassmed.edu/).

TABLE 6.4: ***In silico*** **scoring system for assessment of potential pathogenicity.**

Assessment	***In silico* database/-tool***	***In silico* database/tools result conventions**	**Scoring**	**Scored out of (totalling 10)**
Gene expression in the brain and spinal cord	HBT (brain) GTex (spinal cord)	No expression <6; low expression 6-8; medium-high expression 8+ No expression <5 RPKM; low expression 5-10 RPKM; medium-high expression 10+ RPKM	High-med expression in the brain and spinal cord=2/2, high-med expression in either the brain or spinal cord=1/2, low expression in the brain and SC=1/2, no expression in the brain and SC=0/2	2
In silico protein prediction programs	MutationAssessor MutationTaster Polyphen-2 Pon-P2 SIFT PROVEAN SNPs&GO CADD	Functional (high/medium), non-functional (low/neutral) Disease causing or polymorphism Probably or possibly damaging, or benign Pathogenic, neutral or unknown Damaging or tolerated Deleterious or neutral Disease or neutral Magnitude of rank score (10=top 10% deleterious, 20=top 1% deleterious etc)	Based on the percentage of *in silico* programs returning a potential pathogenic prediction result; <40%=0; 40-60%=0.5, 60-75%=1 , 75-85%=1.5, 85-100%=2	2
Conservation of the affected amino acid across species	Validated protein sequences obtained form NCBI Homologene, and aligned using ClustalOmega PhyloP Phastcons	The identity of the affected residue was compared between humans and all other species, the number of species with the same residue at the corresponding position were considered positive. The total number of species with protein sequence data available, and the percentage of positive species was recorded Conserved residues have positive scores Conserved residues have scores >0.5	**Homologene/ClustalOmega scoring** No. species / 100% / 75-99% / 50-74% / <50% n<4 / 0.6 / 0.4 / 0.2 / 0 n=5-6 / 0.8 / 0.6 / 0.4 / 0 n≥7 / 1 / 0.8 / 0.6 / 0 Conserved = 0.5 Conserved = 0.5	2
Geneic tolerance	RVIS ExAC missence constraint z-score	The RVIS percentile score indicates that the gene is amongst that percentage of most variation intolerant human genes Positive z-scores indicate intolerance to variation	(((100-RVIS%)/100)*2) + (((ExAC constraint z-score)/4) x 2) (If this value is negative score as 0)	4

*Refer to Chapter 2, Section 2.5, Table 2.5 for details of each database/tool.

6.3 Results

6.3.1 *In silico* pipeline for assessment of potential ALS pathogenicity - proof of principle

The *in silico* pipeline and scoring system was applied to 11 known ALS mutations, and three common SNPs to validate its ability to assess potential for ALS pathogenicity, and to determine scoring thresholds. A clear distinction in scores between the two categories was observed, with known ALS mutation scores ranging from 4.80 to 8.01, and common SNP scores ranging from 1.16 to 2.10. Results are summarised in Table 6.5, and detailed in Appendix A.3.4, Table A.3. Analysis of these scores determined that the threshold for a non-synonymous variant to have a high potential for ALS pathogenicity was a score of five, while the threshold for low potential for ALS pathogenicity was a score of two.

Table 6.5: ***In silico*** **assessment of known ALS mutations.**

Variant type	**Gene**	**Amino acid change**	**Gene expression**	**Protein predictons**	**Conservation**	**Genic tolerance**	**Total score (out of 10)**
Known ALS mutation	*SOD1*	p.I114T	2	2	1.8	2.2142	8.0142
Known ALS mutation	*SOD1*	p.E101G	2	1	0.5	2.2142	5.7142
Known ALS mutation	*SOD1*	p.V149G	2	2	1.8	2.2142	8.0142
Known ALS mutation	*FUS*	p.R521C	2	1	1.6	3.1336	7.7336
Known ALS mutation	*FUS*	p.R521H	2	0.5	1.6	3.1336	7.2336
Known ALS mutation	*FUS*	p.R521S	2	1	1.6	3.1336	7.7336
Known ALS mutation	*TARDBP*	p.G294V	2	0	1.6	3.6166	7.2166
Known ALS mutation	*TARDBP*	p.M337V	2	0.5	1.8	3.6166	7.9166
Known ALS mutation	*TARDBP*	p.G376D	2	0	1.4	3.6166	7.0166
Known ALS mutation	*UBQLN2*	p.T487I	2	0	1.4	2.2074	5.6074
Known ALS mutation	*CCNF*	p.S621G	0.5	0.5	1.8	1.9966	4.7966
Common SNP	*TMA16*	p.I176T	1	0	0	0.1566	1.1566
Common SNP	*OR4C3*	p.S100F	0.5	0.5	1.1	0	2.1
Common SNP	*MAP2K3*	p.S39P	0.5	0	0.9	0.6468	2.0468

6.3.2 Novel gene discovery in FALSmq28

WES and WGS datasets were successfully generated for the two affected, and one obligate mutation carrier family members of FALSmq28. SNP array genotyping data was also generated for these three individuals, as well as the 11 additional "at-risk" and two "married-in" family members.

Over 42-fold more variants were detected using WGS compared to WES in FALSmq28. However, performing variant filtering steps substantially reduced this ratio until more coding candidate mutations were identified by WES compared to WGS in Analysis 1 (prior to Sanger sequencing validation). Prior to any variant filtering, WGS also initially detected ~23% more exonic variants (41,774 vs 33,803) including ~21% more non-synonymous variants (18,849 vs 15,528) compared with WES. Figure 6.8 provides a summary of the genomic functional classifications of the variants identified by each sequencing technology across the three members of FALSmq28 (prior to filtering). Interestingly, the vast majority (>99%) of variants detected by WES were shared by all three family members across the three analysis pipelines, however for WGS data this Figure was less than half (~35-40%) (Table 6.6).

The three complementary family-based filtering pipelines that were applied to FALSmq28 WES or WGS data, which considered either the complete exome/genome or genomic regions with LOD scores >0 or >-2 respectively, each produced a distinct list of candidate mutations. Importantly, both autosomal dominant and recessive disease models (and therefore both heterozygous and homozygous variants) were considered in each, as the inheritance pattern in this family was ambiguous. Table 6.6 summarises the sequential reduction of the number of candidate mutations withstanding each filtering step in each of these six pipelines. Importantly, of the 29 candidate mutations identified in FALSmq28 by bioinformatic filtering of NGS data across all six pipelines, just one withstood validation by Sanger sequencing.

Classical family-based analysis of both WES and WGS data failed to identify any validated coding candidate mutations in FALSmq28. Genome-wide linkage analysis of this family subsequently excluded ~86.64% (2,802,514,406bp) of the genome as disease-linked, with LOD scores <-2. Of the remaining ~13.36% (432,315,594bp) of the genome possibly linked to disease in this family, just ~2.26% (73,123,343bp) showed LOD scores >0. The highest LOD score of 1.1924 was found on chromosome 18. Figure 6.9 illustrates the LOD score results obtained from linkage analysis across the genome.

When considering all genomic regions other than intergenic or intronic, family-based analysis identified just one validated homozygous candidate mutation, upstream of the gene *MIR512*, which fell within a non-excluded linkage region (Analysis 3). The details of this genomic variant are provided in Table 6.7. However, Sanger sequencing of seven non-related controls showed this variant was also present in five of these control individuals, in both a heterozygous (n=3) and homozygous (n=2) state (Figure 6.10).

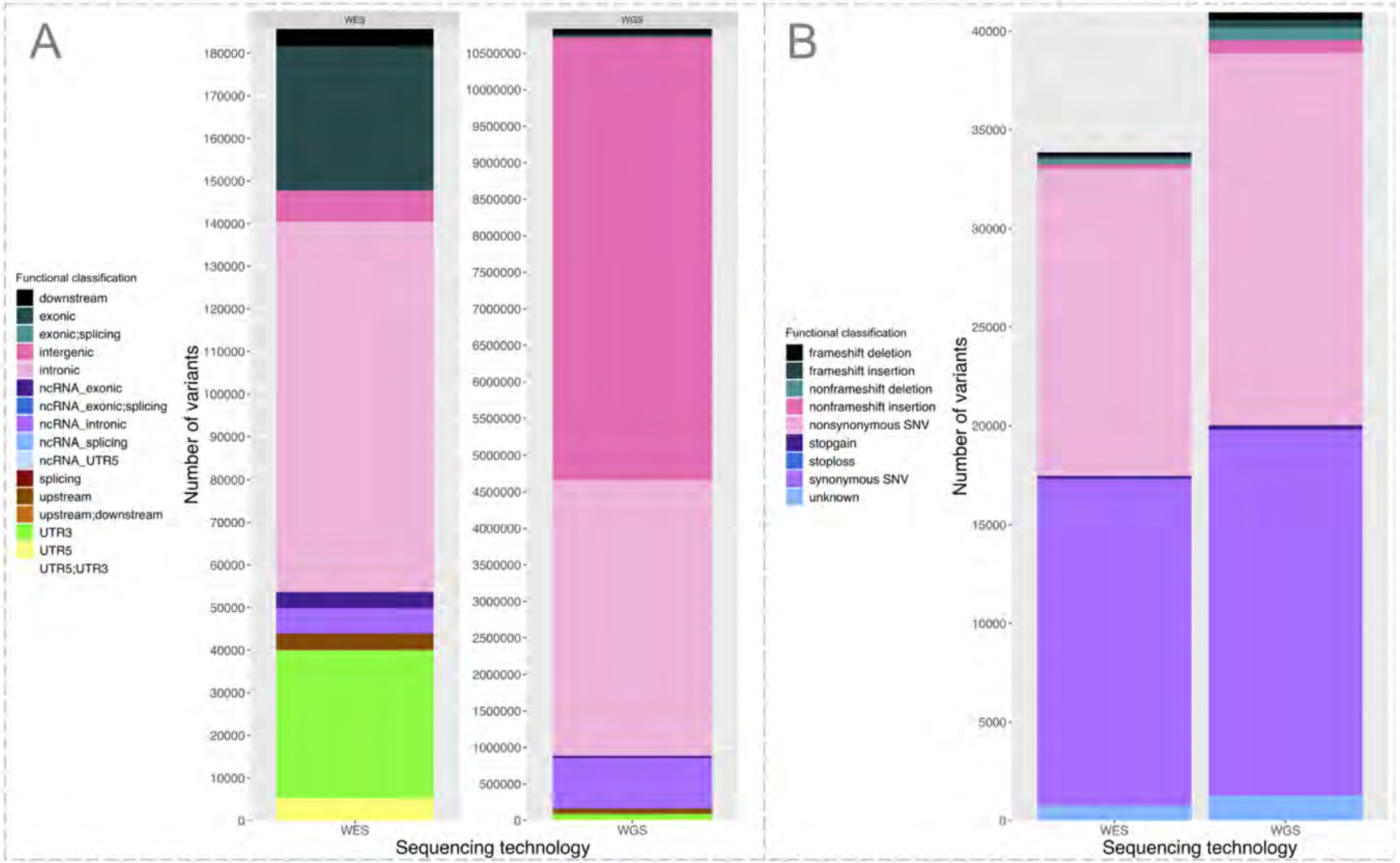


Figure 6.8: **Stacked bar charts showing the distribution of FALSmq28 WES and WGS variants across genomic functional classes.** (A) Variant distribution between the major genomic functional classes. (B) Exonic variant distribution between the exonic functional classes.

Table 6.6: **Filtering results of family-based analysis of family FALSmq28.**

Step	**Description of remaining variants**	**Number of variants remaining**					
		WES			**WGS**		
		Analysis 1 Traditional	Analysis 2 LOD>0	Analysis 3 LOD>-2	Analysis 1 Traditional	Analysis 2 LOD>0	Analysis 3 LOD>-2
Family VCF	Total variants across all family members	183,991			7,799,575		
Linkage analysis	Potentially disease-linked	.	5,478	28,130	.	184,908	995,846
Quality filtering	GQ>20	.	.	10,143	.	.	821,052
Custom family analysis	Shared variants	182,408	5,448	10,129	2,792,679	66,371	341,055
First-tier filtering	Absent from dbSNP147 and/or gnomAD NFE MAF<0.0001	3,252	119	59	413,262	10,091	41,481
	Coding (and regulatory)	229	43	16	278	287	2,763
	Amino acid (and regulatory element) altering	170	.	.	255	.	.
Special filtering	MGRB AC<2	.	.	.	.	.	134
	Variants filtered in analysis 1 or 2 removed	.	.	10	.	.	109
First-tier validation	Correct variant calls	30	11	2	210	235	100
Second-tier filtering	gnomAD and ExAC AC<2	20	8	0	19	22	12
	MGRB and Project MiNE controls AC<2	15	1	0	3	2	9
Second-tier validation	Validated by Sanger sequencing	0	0	0	0	0	1

Abbreviations: GQ, genotype quality; Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MAF, minor allele frequency; MGRB, Medical Genome Reference Bank; and AC, allele count.

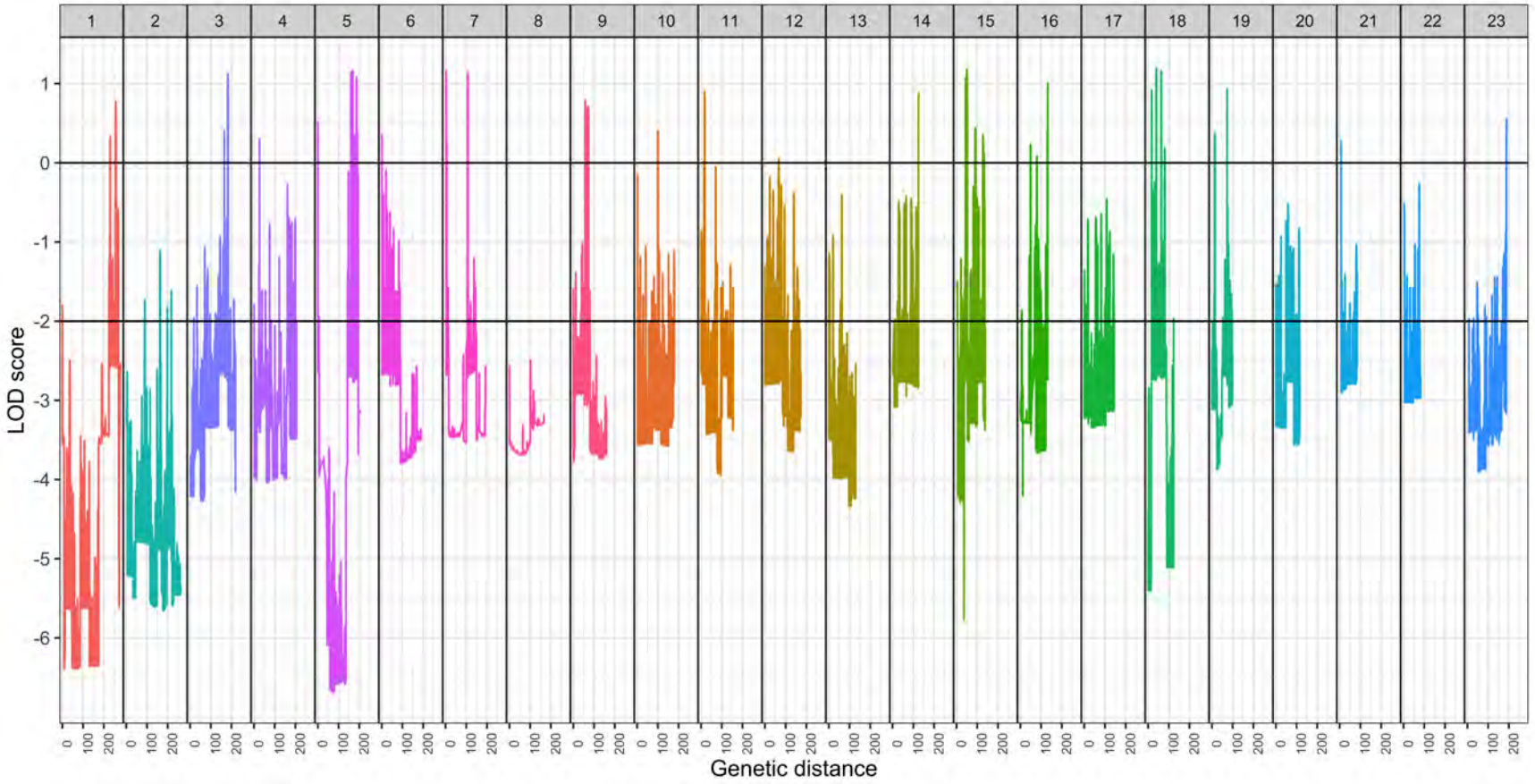


Figure 6.9: **Results of genetic linkage analysis of FALSmq28.** The genetic distance across each chromosome is presented on the x-axis, and LOD scores are shown on the y-axis. The numbers in the grey boxes indicate the relevant chromosome. Each chromosome is represented by a different line colour. Peaks falling below LOD -2 represent genomic regions excluded as disease-linked.

TABLE 6.7: **Details of the FALSmq28 candidate mutation.**

Gene.refGene	*MIR512-1*;*MIR512-2*
Func.refGene	upstream
gDNA alteration	g.19:54169255C>A
Family genotype	homozygous
ExAC	absent
gnomAD	38 heterozygotes of African descent
DACC	absent
MGRB	absent
Australian controls - Sanger sequencing	two homozygotes three heterozygotes two wild-type

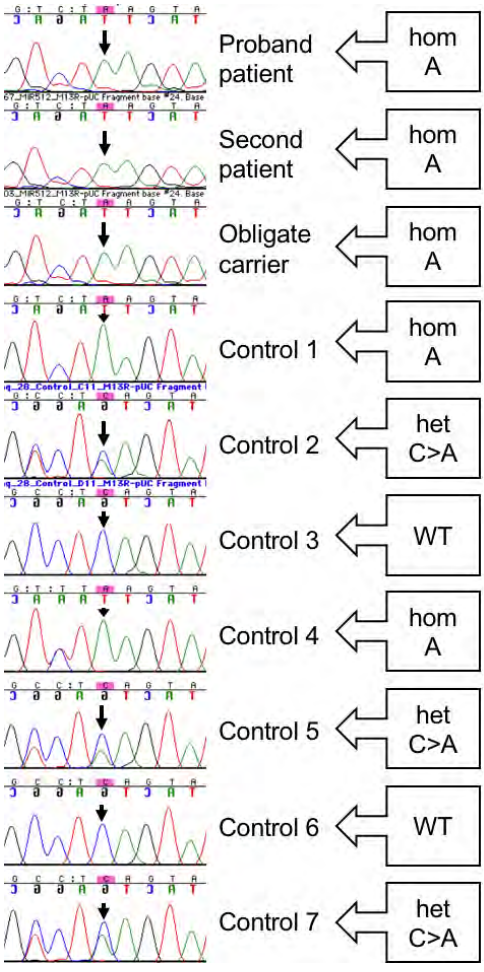


FIGURE 6.10: **Chromatograms of the FALSmq28 *MIR512* candidate mutation.** The FALSmq28 homozygous candidate mutation *MIR512* g.19:54169255C>A identified in WGS data was screened through all three informative FALSmq28 family memebers and seven unrelated control individuals. The arrows indicate the corresponding nucleotide base in each individual.

6.3.3 Novel gene discovery in small ALS families

The sequential reduction in the number of variants withstanding family-based filtering and validation in FALS15, FALS45, FALSmq2 and FALSmq20 are summarised in a step-wise manner in Table 6.8. Bioinformatics filtering in each family narrowed the search for their disease causal mutation to less than 0.7% of their exome variants. Custom family analysis found an average of 48.09% of variants to be candidate causal mutations, with common variant filtering removing an average of 98.69% of these from analysis. Among the remaining candidate variants, an average of 25.79% fell in protein-coding regions, though only 61.84% of these altered the amino acid sequence. Alarmingly, 40.68% of these had an incorrect alternate allele call from bioinformatics processing, and were thus removed as candidate variants. Refined filtering for less common population-based SNPs reduced the number of variants by a further 43.45%. Finally, 6.83% of bioinformatically filtered variants were found to be sequencing artefacts upon Sanger sequencing validation. Due to cost and time constraints, only the top ten FALSmq20 candidate mutations (according to *in silico* scoring as described below) were validated by Sanger sequencing, all of which were confirmed in both affected individuals from this family. Using updated control databases for common variant filtering, one candidate variant was found to be a rare benign SNP, and was removed from analysis. Altogether, this analysis identified a total of 20, 11, 16 and 64 candidate mutations in each of FALS15, FALS45, FALSmq2 and FALSmq20, respectively. The remaining candidate mutations for each family are summarised in Tables 6.9, 6.10, 6.11 and 6.12.

Of these candidate mutations, five, six, one and 11 from each family were assessed as having a high potential pathogenicity using the *in silico* pipeline developed here. Detailed results of this *in silico* assessment of pathogenicity are presented in Appendix A.3.4, Tables A.5, A.6, A.7 and A.8. Figures 6.11 and 6.12 show examples of the graphical outputs obtained from the Human Brain Transcriptome (HBT; `http://hbatlas.org/`; Kang et al., 2011; Pletikos et al., 2014) (used to assess gene expression in the brain) and multiple protein sequence alignment using NCBI homologene (`http://www.ncbi.nlm.nih.gov/homologene`) and Clustal Omega v1.2.4 (`http://www.ebi.ac.uk/Tools/msa/clustalo`; Sievers et al., 2011) (used to assess amino acid conservation across species), respectively. The supportive *in silico* data collected for each remaining candidate mutation is presented in Appendix A.3.5, Tables A.9, A.10, A.11 and A.12.

TABLE 6.8: **Filtering results of family-based analysis of families FALS15, FALS45, FALSmq2 and FALSmq20.**

Step	**Description of remaining variants**	**Number of variants remaining**			
		FALS15	**FALS45**	**FALSmq2**	**FALSmq20**
Family VCF	Total variants across family members	90,418	95,285	97,409	93,065
Custom family analysis	Shared variants	55,583	16,384	62,298	46,280
First-tier filtering	Absent from dbSNP129, dbSNP142, 1000Genomes and/or ExAC NFE MAF<0.0001	465	453	376	479
	Coding	103	84	99	173
	Amino acid altering	52	55	66	112
First-tier validation	Correct variant calls	27	32	35	83
Second-tier filtering	ExAC NFE AC<2	25	25	25	74
	ExAC NFE, DACC and MGRB AC<2	21	25	22	74
	ExAC NFE, DACC, MGRB and gnomAD NFE AC<2	20	14	17	64
Second-tier validation	Validated	20	11	16	64
Updated filtering	MGRB and Project MiNE controls AC<2	19	11	16	64*

Abbreviations: Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MAF, minor allele frequency; MGRB, Medical Genome Reference Bank; and AC, allele count.
*Only top 10 candidate mutations from FALSmq20 underwent PCR and Sanger sequencing validation.

Table 6.9: **FAL15 candidate mutations.**

Gene	Accession number	cDNA change	Amino acid change	Control databases*	Presence in other ALS cohorts*	Total score (out of 10)	Priority category	Priority Ranking
CLCN4	NM_001830	c.T2003C	p.I668T	absent	absent	8	high	1
MTSS1L	NM_138383	c.G376A	p.A126T	gnomAD/ExAC (AC=1 NFE)	Project MinE - SALS (AC=1)	6.8546	high	2
SCN4A	NM_000334	c.C673T	p.R225W	gnomAD (FILTERED; AC=6; 4 NFE)	absent	6.3414	high	3
LRRN2	NM_006338	c.T587C	p.I196T	absent	absent	6.0844	high	4
SUPV3L1	NM_003171	c.C502G	p.Q168E	absent	absent	5.683	high	5
HOXD3	NM_006898	c.A746G	p.Y249C	gnomAD (FILTERED; AC=1 other)	absent	4.7454	medium	6
FAM171A1	NM_001010924	c.A1553G	p.H518R	gnomAD/ExAC (AC=1 NFE)	absent	4.5454	medium	7
SP1	NM_003109	c.G433A	p.A145T	gnomAD (AC=1 NFE)	absent	4.089	medium	8
MAPKAPK3	NM_004635	c.A1103G	p.K368R	absent	absent	3.98	medium	9
SIM1	NM_005068	c.G2198T	p.G733V	absent	absent	3.8374	medium	10
ZNF385B	NM_152520	c.C1303T	p.P435S	absent	absent	3.6654	medium	11
TYMP	NM_001953	c.C733G	p.Q245E	absent	absent	2.935	medium	12
TNS2	NM_015319	c.C2975T	p.S992L	gnomAD (AC=1 NFE)	absent	2.5	medium	13
NECAB3	NM_031231	c.G608T	p.R203L	absent	absent	2.485	medium	14
ZNF425	NM_001001661	c.G1271C	p.R424P	absent	absent	1.9028	low	15
CEP295	NM_033395	c.A5120T	p.N1707I	absent	absent	1	low	16
ZNF497	NM_198458	c.A68G	p.K23R	absent	absent	0.6	low	17
ZNF497	NM_198458	c.T65G	p.V22G	absent	absent	0.6	low	18
RNF133	NM_139175	c.G281A	p.R94Q	gnomAD (AC=1 SEA)	absent	0	low	19

Abbreviations: Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MGRB, Medical Genome Reference Bank; and AC, allele count.

Table 6.10: **FAL45 candidate mutations.**

Gene	Accession number	cDNA change	Amino acid change	Control databases*	Other ALS cohorts*	Total score (out of 10)	Priority category	Priority Ranking
SCCPDH	NM_016002	c.G766T	p.V256L	gnomAD/ExAC (AC=2 NFE)	Project MinE - SALS (AC=1)	6.7012	high	1
GDPD1	NM_182569	c.C661A	p.P221T	gnomAD (AC=1 NFE)	850 WGS VCF - SOD1 FALS (AC=1)	6.1664	high	2
SPATA2	NM_006038	c.G616A	p.G206S	gnomAD/ExAC (AC=5; 2 NFE)	absent	5.778	high	3
KRT85	NM_002283	c.T13C	p.S5P	gnomAD (AC=1 NFE)	850 WGS VCF - SALS (AC=1)	5.7774	high	4
GABRG3	NM_033223	c.C707T	p.S236F	absent	absent	5.3252	high	5
GRIN2D	NM_000836	c.G430T	p.V144L	absent	absent	5.295	high	6
HIST1H3G	NM_003534	c.C115T	p.P39S	gnomAD/ExAC (AC=3; 1 NFE)	absent	4.7026	medium	7
PIGZ	NM_025163	c.T180A	p.D60E	gnomAD (AC=1 NFE); MGRB (AC=1)	absent	4.3976	medium	8
NPBWR1	NM_005285	c.C754G	p.L252V	gnomAD/ExAC (AC=1 NFE)	absent	3.4708	medium	9
ORM1	NM_000607	c.G414T	p.K138N	absent	absent	1.9974	low	10
ZNF132	NM_003433	c.G1363A	p.G455R	gnomAD/ExAC (AC=5; 1 NFE)	absent	1.7342	low	11

Abbreviations: Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MGRB, Medical Genome Reference Bank; and AC, allele count.

Table 6.11: **FALmq2 candidate mutations.**

Gene	Accession number	cDNA change	Amino acid change	Control databases*	Presence in other ALS cohorts*	Total score (out of 10)	Priority category	Priority Ranking
STRN4	NM_013403	c.T1086A	p.D362E	MiNE controls (AC=1); ExAC (AC=1 NFE)	absent	5.8334	high	1
EHBP1	NM_001142614	c.A1856T	p.Q619L	gnomAD (AC=1 Latino)	absent	4.2856	medium	2
ZFHX2	NM_033400	c.1694_1695delCCinsGAC	p.T565Rfs*19	absent	absent	4.1	medium	3
CHRNA2	NM_000742	c.G1231C	p.E411Q	gnomAD (FILTERED; AC=1 NFE)	absent	3.2016	medium	4
TUSC5	NM_172367	c.G424A	p.A142T	gonmAD/ExAC (AC=3; 2NFE)	absent	2.6	medium	5
EMP2	NM_001424	c.T368G	p.I123S	gnomAD (AC=1 NFE)	absent	2.4064	medium	6
DPH6	NM_080650	c.A655G	p.I219V	absent	absent	2.325	medium	7
ALPK1	NM_025144	c.G2935A	p.D979N	gnomAD /ExAC (AC=1 NFE)	absent	2.2	medium	8
P2RY2	NM_002564	c.T46C	p.W16R	absent	absent	1.8528	low	9
SLC25A21	NM_030631	c.C442T	p.P148S	DACC (AC=1)	absent	1.6484	low	10
PCDHB11	NM_018931	c.T2275A	p.S759T	absent	absent	1.514	low	11
CFH	NM_000186	c.C1262G	p.A421G	absent	absent	1.1726	low	12
FANCC	NM_000136	c.C591G	p.D197E	gnomAD (AC=1 NFE)	850 WGS VCF - SALS (AC=1)	1	low	13
ANKRD18B	NM_001244752	c.T1766G	p.L589R	absent	absent	0.6	low	14
CFAP47	NM_173695	c.G96T	p.Q32H	gnomAD/ExAC (AC=1 NFE)	absent	0.5	low	15
CFAP47	NM_173695	c.G97C	p.D33H	gnomAD/ExAC (AC=1 NFE)	absent	0.5	low	16

Abbreviations: Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MGRB, Medical Genome Reference Bank; and AC, allele count.

Table 6.12: **FALmq20 candidate mutations.**

Gene	Accession number	cDNA change	Amino acid change	Control databases*	Presence in other ALS cohorts*	Total score (out of 10)	Priority category	Priority Ranking
RASGRF1	NM_002891	c.C101G	p.S34W	absent	absent	9.4	high	1
NCOR2	NM_006312	c.G6437A	p.R2146Q	gnomAD (AC=2 African)	absent	8.6286	high	2
TAZ	NM_000116	c.C29G	p.P10R	absent	absent	7.2258	high	3
HIC2	NM_015094	c.C1577T	p.T526M	gnomAD/ExAC (AC=9; 1 NFE)	Project MiNE (AC=1); ALSdb (AC=1)	6.403	high	4
CRIM1	NM_016441	c.G2980A	p.G994R	absent	absent	5.8614	high	5
SLC35A4	NM_080670	c.T853C	p.C285R	gnomAD/ExAC (AC=1 NFE)	absent	5.6842	high	6
ELFN2	NM_052906	c.C907T	p.H303Y	gnomAD (AC=1 NFE)	absent	5.5134	high	7
POU2F2	NM_002698	c.C245T	p.P82L	gnomAD/ExAC (AC=4; 2 NFE)	absent	5.4848	high	8
DNAJC4	NM_005528	c.C292T	p.Q98X	absent	absent	5.4694	high	9
NUDC	NM_006600	c.G609C	p.Q203H	absent	absent	5.386	high	10
SLC24A2	NM_020344	c.G856A	p.V286I	absent	absent	5.0904	high	11
MAP1A	NM_002373	c.A1082G	p.K361R	absent	absent	4.8578	medium	12
CSMD3	NM_052900	c.C7415G	p.P2472R	gnomAD (AC=2; 0 NFE)	AVS SALS (AC=1)	4.793	medium	13
OPRK1	NM_000912	c.G542C	p.C181S	absent	absent	4.646	medium	14
SOX15	NM_006942	c.G356A	p.R119Q	gnomAD/ExAC (FILTERED; AC=2; 1 NFE); MGRB (AC=1)	absent	4.55	medium	15
COL3A1	NM_000090	c.C2638A	p.L880I	absent	absent	4.4482	medium	16
TSN	NM_001261401	c.C394T	p.R132C	gnomAD (AC=1 Latino)	absent	4.3992	medium	17
SARAF	NM_016127	c.A127G	p.K43E	gnomAD (AC=1 NFE)	absent	4.3794	medium	18
MIEF1	NM_019008	c.C107T	p.A36V	gnomAD (AC=1 other)	absent	4.3568	medium	19
TMEM199	NM_152464	c.C41G	p.A14G	absent	absent	4.2592	medium	20

ENPP5	NM_021572	c.C725T	p.T242M	gnomAD/ExAC (AC=5; 2 NFE)	absent	4.1986	medium	21
TSSK4	NM_174944	c.G613A	p.A205T	gnomAD (AC=1 NFE); Project MiNE controls (AC=1)	absent	4.196	medium	22
SLC7A14	NM_020949	c.G2152A	p.E718K	absent	absent	4.1402	medium	23
SLC7A14	NM_020949	c.A2144G	p.E715G	absent	absent	4.1402	medium	24
TTN	NM_003319	c.G55441A	p.A18481T	gnomAD/ExAC (AC=1 NFE)	absent	4	medium	25
LHX1	NM_005568	c.C970G	p.L324V	gnomAD/ExAC (AC=1 NFE)	absent	3.9798	medium	26
E2F8	NM_024680	c.A293C	p.H98P	Project MiNE controls (AC=1)	absent	3.933	medium	27
ECE2	NM_014693	c.G580A	p.G194S	absent	absent	3.5382	medium	28
IRX6	NM_024335	c.A383G	p.E128G	gnomAD (AC=1 other)	absent	3.4784	medium	29
REG3G	NM_198448	c.T311A	p.I104N	absent	Project MiNE (AC=1)	3.4	medium	30
ALDH3B1	unknown	g.chr11:67786087A>G	unknown	gnomAD (AC=1 NFE)	absent	3.295	medium	31
DNAJC13	NM_015268	c.A3829G	p.K1277E	absent	Project MiNE (AC=1)	3.2046	medium	32
MARVELD2	NM_001038603	c.A1361G	p.E454G	absent	absent	3.0764	medium	33
BAHCC1	unknown	g.chr17:79428542G>A	unknown	gnomAD/ExAC (FILTERED; AC=5 NFE)	absent	3	medium	34
USP53	NM_019050	c.G1817A	p.S606N	absent	absent	2.8878	medium	35
RDH12	NM_152443	c.C112T	p.P38S	gnomAD (AC=1 SEA)	absent	2.8512	medium	36
OR4Q3	NM_172194	c.T134C	p.I45T	absent	absent	2.8482	medium	37
OR2K2	NM_205859	c.C713T	p.S238F	gnomAD/ExAC (AC=2 NFE)	absent	2.8	medium	38
RSRP1	NM_020317	c.C809G	p.A270G	absent	absent	2.7962	medium	39
ABHD15	NM_198147	c.G121A	p.A41T	absent	absent	2.7086	medium	40
FCHSD1	NM_033449	c.G2012A	p.R671H	gnomAD/ExAC (AC=3; 1 NFE)	absent	2.6492	medium	41
OR4D9	NM_001004711	c.T220C	p.S74P	absent	absent	2.1	medium	42

HR	NM_005144	c.G1745A	p.R582Q	gnomAD (AC=2 NFE)	absent	1.9738	high	43
KIF26A	NM_015656	c.G181A	p.G61S	gnomAD (AC=1 NFE)	Project MiNE (AC=1)	1.6	low	44
PNMAL2	NM_020709	c.A676T	p.T226S	absent	absent	1.6	low	45
LMTK3	NM_001080434	c.C1922T	p.P641L	gnomAD (AC=2 African)	absent	1.6	low	46
FCGBP	NM_003890	c.C11948T	p.P3983L	gnomAD/ExAC (AC=6; 2 NFE)	absent	1.5	low	47
MGAM	NM_004668	c.T2389C	p.W797R	gnomAD/ExAC (AC=2 SEA)	absent	1.5	low	48
MRPS28	NM_014018	c.A191C	p.Q64P	gnomAD (AC=5; 1 NFE)	absent	1.3072	low	49
ERVV-1	NM_152473	c.A1169T	p.Y390F	absent	absent	1.1	low	50
MROH5	unknown	g.chr8:142480784G>C	unknown	gnomAD/ExAC (AC=3; 1 NFE)	absent	1.0052	low	51
SLC22A24	NM_001136506	c.G1154A	p.C385Y	gnomAD (AC=1 Latino)	absent	1	low	52
MXRA5	NM_015419	c.G7747A	p.D2583N	gnomAD/ExAC (AC=4; 2 NFE)	absent	1	low	53
CGREF1	NM_006569	c.G937A	p.V313M	gnomAD/ExAC (AC=3; 2 NFE)	absent	0.889	low	54
FASTKD2	NM_014929	c.G458A	p.R153H	gnomAD/ExAC (AC=8; 2 NFE)	absent	0.724	low	55
PLEKHG4B	NM_052909	c.G1615T	p.A539S	absent	absent	0.6778	low	56
FANCA	NM_000135	c.T3901C	p.S1301P	gnomAD/ExAC (AC=3; 0 NFE); Project MiNE controls (AC=1)	Project MiNE (AC=1)	0.6	low	57
LOC79999	NM_001291904	c.A192C	p.Q64H	absent	absent	0.5	low	58
DNAH11	NM_001277115	c.C5375T	p.P1792L	Project MiNE controls (AC=1)	absent	0.5	low	59
PLEKHG4B	NM_052909	c.G1510A	p.V504M	gnomAD/ExAC (AC=17; 2 NFE)	absent	0.4778	low	60
LIPF	NM_004190	c.G79A	p.G27R	absent	absent	0.2798	low	61
MUC16	NM_024690	c.G7073A	p.R2358Q	gnomAD (AC=1 other)	absent	0.2	low	62
CEP295	NM_033395	c.A4471G	p.K1491E	gnomAD (AC=1 NFE)	absent	0	low	63
KRTAP29-1	NM_001257309	c.T410A	p.M137K	absent	absent	0	low	64

Abbreviations: Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; NFE, Non-Finnish Europeans; MGRB, Medical Genome Reference Bank; and AC, allele count.

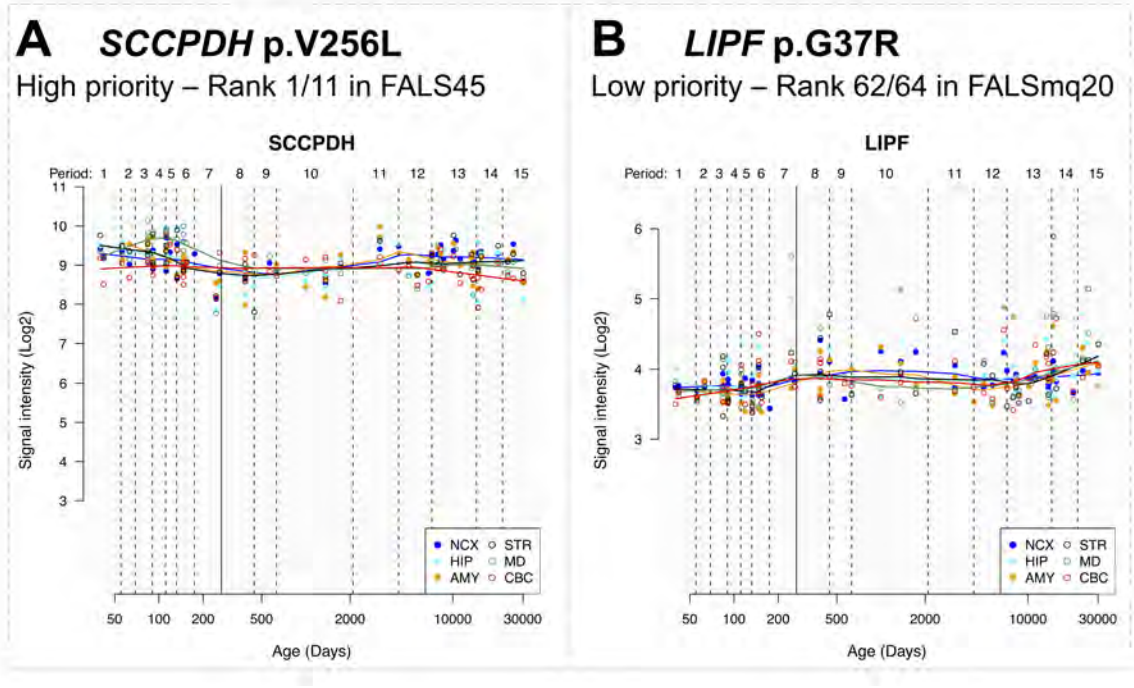


Figure 6.11: **Examples of HBT gene expression graphs used in the *in silico* pipeline for assessment of potential ALS pathogenicity.** (A) Example of HBT gene expression graph for the high priority candidate mutation *SCCPDH* p.V256L, ranked first of 11 in FALS45, showing high gene expression. (B) Example of HBT gene expression graph for the low priority candidate mutation *LIPF* p.G37R, ranked 62 of 64 in FALSmq20, showing low gene expression.

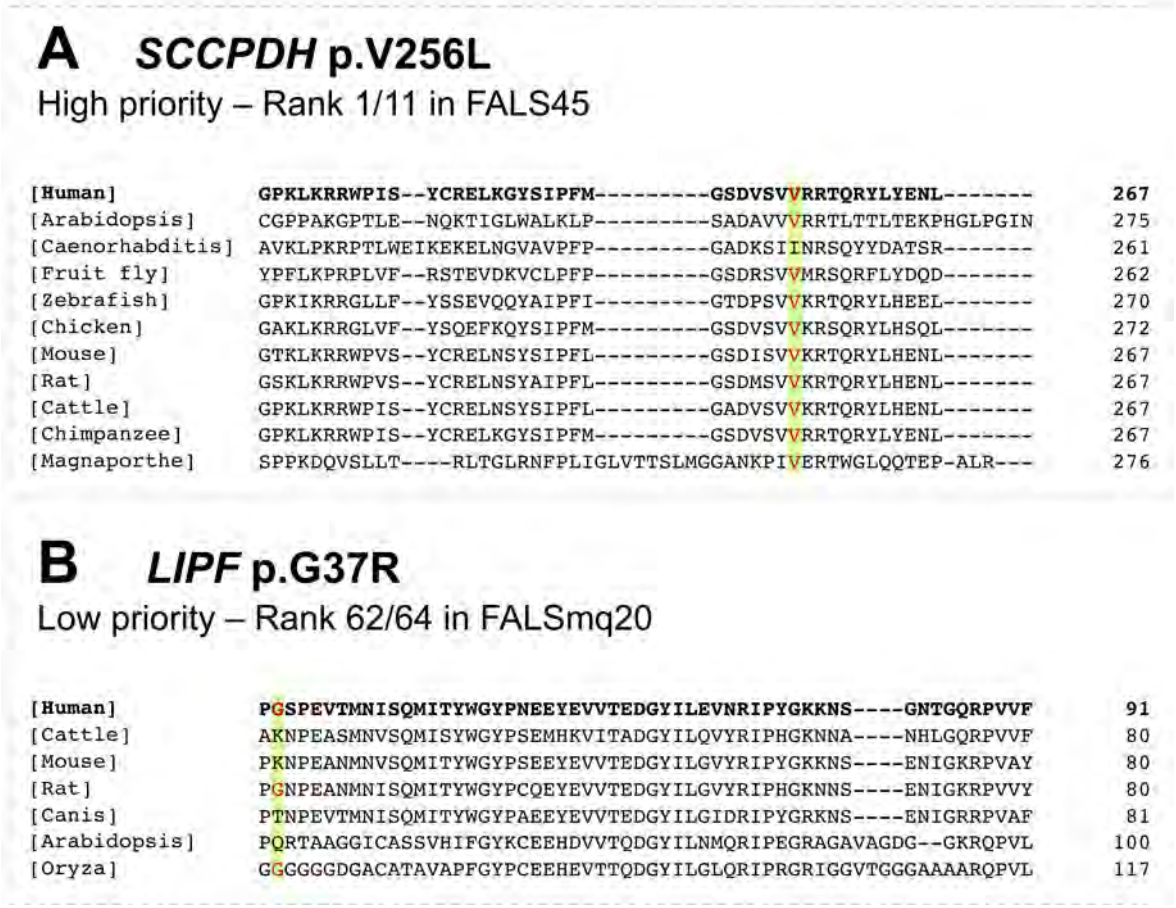


FIGURE 6.12: **Examples of multiple sequence alignment used in the *in silico* pipeline to assess potential pathogenicity.** (A) Example of multiple sequence alignment for the high priority candidate mutation *SCCPDH* p.V256L, ranked first of 11 in FALS45, showing high species conservation. (B) Example of multiple sequence alignment for the low priority candidate mutation *LIPF* p.G37R, ranked 62 of 64 in FALSmq20, showing low species conservation. The residue substituted for by a candidate mutation is highlighted in green. Red denotes a residue matching the wild-type human residue.

6.4 Discussion

In this Chapter, a list of candidate gene mutations have been identified for each of five Australian ALS families that are negative for all known ALS genes. This was achieved by employing custom family-based analysis pipelines utilising combinations of WES, WGS, bioinformatics, genetic linkage and validation strategies. For small families that have many candidate mutations after genetic filtering, an *in silico* pipeline was also developed to assess the potential ALS pathogenicity of those protein-altering candidate mutations using various tools and databases. This pipeline was successfully applied to prioritise and rank the list of candidate mutations from each small family. Family-based analysis of WES data from the families FALS15, FALS45, FALSmq2 and FALSmq20 identified 19, 11, 16 and 64 candidate mutations, respectively, from an initial ~80,000 variants in each family. Application of the *in silico* pipeline implicated just five, six, one and 11 of these as having a high potential for ALS pathogenicity, one of which is likely to be causing ALS in each family. Top scoring variants are suitable for downstream *in vitro* studies to further elucidate their potential contribution to ALS pathogenesis. While no single nucleotide level candidate mutation was identified in FALSmq28, genome-wide linkage analysis substantially narrowed the search for the ALS causal mutation to less than 14% of the genome, highlighting the immense benefits of including this technique in family-based WGS studies.

6.4.1 Novel gene discovery in ALS

As described previously, the major genetic discoveries in ALS have resulted from combining the power of genome-wide linkage analysis and NGS. Our research group has successfully applied this approach (including the bioinformatic pipelines described in this Chapter) to identify novel ALS mutations in *TARDBP* (Sreedharan et al., 2008), *UBQLN2* (Williams et al., 2012b), *TBK1* (Williams et al., 2015) and *CCNF* (Williams et al., 2016b). As such, the cause of disease in all large Australian ALS families has been identified. However, this leaves the majority of the smaller, and therefore more challenging, families to be solved. The five families analysed in this Chapter represent typical such families. These families possess various characteristics which hinder traditional disease gene mapping approaches (as established in Chapter 1, Section 1.6.1). This is primarily due to limited sample availability, caused by the late and highly variable age of disease onset of ALS, and/or reduced penetrance of causal mutations. Furthermore, the high degree of genetic heterogeneity in ALS, and the fact that many causal mutations are rare, dictates that each family must be

considered in isolation, because searching for mutations (or genes) shared in multiple families may discard causal mutations. This also extends to linkage analysis, as results from multiple families cannot be combined unless they carry mutations in the same gene. The low incidence of disease, particularly the familial form, also complicates replication efforts. Indeed, it is possible that some of the unidentified ALS mutations are in fact private mutations, largely restricted to a single family. To establish the causality of such rare mutations using genetics alone would be exceedingly difficult.

6.4.2 The NGS family-based pipeline

The core of the gene discovery pipeline applied here was the bioinformatic filtering of NGS data generated from the informative members of each ALS family, to reduce the number of candidate mutations. The pivotal steps of this pipeline involved custom family-based segregation analysis (ie. identifying variants identical by descent), bioinformatics filtration of population-based variants and variant validation. Each of these steps plays a critical role in disease gene discovery. The intricacies of segregation analysis and common variant filtering are discussed in the following sections, while variant validation will be discussed in Chapter 8, Section 8.3.3.

6.4.2.1 Segregation analysis

Segregation analysis using NGS data formed the basis for novel gene discovery in all five ALS families analysed in this Chapter. To identify variants identical by descent, Custom Scripts were applied to WES or WGS data to extract all nucleotide level variants which were shared by all affected (or obligate mutation carrier) family members and absent from any "married-in" control individuals. This is based on the principle that a rare pathogenic genetic mutation is being inherited within each family. Unfortunately, given the limited sample availability in the ALS families analysed here, the power of segregation analysis was diminished. Additional affected family members, or informative "married-in" controls, would allow higher order comparisons to increase stringency and reduce the number of shared variants (or, variants identical by descent). The close relationships between the sequenced family members also contributed to the large number of shared variants identified. In each of FALS15, FALS45, FALSmq2 and FALSmq20, the two available family members were first degree parent-offspring pairs, meaning each pair shared 50% of their DNA sequence. In contrast, the two ALS patients in FALSmq28 were second cousins, sharing an

average of just 3.13% of their genetic information. While we would expect that this would lead to a far smaller number of shared variants in FALSmq28 compared to the parent-offspring pair families, this is not reflected in our results. However, the inflation in the number of shared variants in FALSmq28 is attributable to our consideration of both autosomal dominant and recessive disease models, rather than just the autosomal dominant model adopted in all other families.

Various bioinformatics tools are available to perform shared variant analysis. These include commands in the BCFtools (Li, 2011) and VCFtools (Danecek et al., 2011) programs, as well as commercial programs such as Ingenuity Variant Analysis (Qiagen). Here, custom bioinformatics scripts were written using R. This was the preferred method as it allowed a great deal of customisation, and also enabled data visualisation. Firstly, this allowed the same basic scripting strategy for WES data from the small families to be applied to both WES and WGS data analysis in FALSmq28. R scripting also allowed each filtering step to be completed separately, so that the reduction in variant numbers could be attributed to each specific filter. The flexibility also facilitated different combinations of filtering steps to be performed seamlessly. By using the RStudio graphical interface, it was possible to visually observe the effect of each filtering step on the number of variants present in the VCF.

6.4.2.2 Common variant filtering

As was established in Chapter 2, Section 2.4.3, rigorous control filtering is a crucial step for the removal of benign variants from analysis, and establishing the novelty of candidate mutations. A benign classification for a genetic variant is strongly supported by an allele frequency in the control population which is greater than that expected for the disease mutation, while an allele frequency greater than 5% is considered standalone support for a benign classification (Richards et al., 2015). In this Chapter, various control databases were utilised for control variant filtering. There were some differences between the control databases utilised for filtering of the small families compared with FALSmq28. This was necessary as FALSmq28 underwent NGS data generation, bioinformatic processing and annotation at a separate and later stage than the other families. With the continual expansion of control databases to include more individuals, updated filtering was also required.

Common variants were defined as those which were present in one of the well-established control databases of dbSNP (including versions dbSNP129, dbSNP142

and dbSNP147) (Sherry et al., 2001) or 1000Genomes (Auton et al., 2015). While there is considerable overlap between the genetic variants reported in these databases, a substantial number of low frequency variants are reported in just one database. This is likely attributable to sample size and/or filtering criteria. As such, it is necessary to utilise the catalogue of variation from the combination of these control databases to comprehensively filter for non-damaging genetic variants. The dbSNP database is a central repository for small genetic variants (SNPs or small indels), which are each reviewed based on a number of evidence criteria. The 1000Genomes database contains a catalogue of genomic variants assessed to have a MAF>0.01 among 1,092 healthy individuals. These databases are widely accepted to represent the catalogue of common benign small genetic variants. VCFs are commonly annotated to include whether a variant is reported by each of these control databases, as was achieved here using ANNOVAR. As such, a filter was employed to remove any variants present in these databases as part of the novel gene discovery pipeline described here, with the intention of removing population-based benign genetic variants. However, this approach is not perfect, as these control databases are known to contain rare pathogenic variants. For instance, the ALS mutation FUS p.N63S is reported by both dbSNP134 (as well as all subsequent dbSNP releases) and 1000Genomes. However, these filters were necessary to reduce the number of variants under analysis to a manageable number. Additionally, the known ALS mutations present in these databases, as well as all known ALS genes, had already been screened in these families, and the likelihood of removing any novel pathogenic variants was low.

In addition to these more established databases, two more recently curated international control databases were interrogated, being ExAC and gnomAD (Table 2.4). The ExAC database is an aggregate of WES data from 60,706 healthy, unrelated individuals sequenced as part of a variety of case-control and population studies. gnomAD is an expansion of ExAC, containing the majority of the WES data from its predecessor as well as additional WES and WGS data, to total 123,136 WES and 15,496 WGS sequences. The ExAC and gnomAD databases also contain a number of pathogenic mutations, including the ALS mutations *SOD1* p.I114T and *TARDBP* p.M337V. Fortunately, when utilising these the ExAC and gnomAD databases in the filtering pipeline, the number of variants under analysis had reached a manageable number. Therefore, rather than simply filtering variants based on membership to these databases, a more conservative approach utilising MAF and allele count thresholds was applied. The MAF threshold of 0.0001 was set based upon the frequency with which the aforementioned *SOD1* and *TARDBP* mutations were observed in the ExAC

and gnomAD databases. The allele count threshold was set at two for this same reason.

Most pathogenic mutations found within these control databases cause late onset diseases. Participants included in these databases are labelled as "healthy control individuals" as they are disease-free at the time of recruitment. However, in the absence of follow-up clinical consultations, it is never known whether any of these individuals go on to develop disease later in life. Therefore, it is necessary to exercise caution with any variants reported in only one or two database individuals. Adding to this, clinical data is not readily available from participants for ethical reasons. Therefore, individuals may be included in a sub-study as a control, but may not be an appropriate control for our purposes. For example, an individual may have cognitive decline, and while this would not exclude them as a control for diabetes research, they would not ordinarily be included as a control in studies of ALS. Further, most variants reported in these databases have not been validated by Sanger sequencing, therefore some may be sequencing artefacts. Together, this reinforces the necessity for conservative approaches to common variant filtering to avoid the removal of pathogenic mutations.

Population-stratification is another important consideration in variant filtering. As established in Chapter 1, Section 1.4, particular ALS mutations cluster in patients of certain ancestral backgrounds. Additionally, the drastic effect of using population-matched controls was discussed in Chapter 5. Therefore, in addition to the international control databases, two control databases of healthy Australians, DACC and MGRB (detailed in Table 2.4), were also utilised. Importantly, when utilising gnomAD and ExAC, filtering was based on non-Finnish Europeans (NFE) control individuals, as this cohort has the most similar ancestral background to the Australian families analysed in this Chapter.

6.4.2.3 Genome-wide linkage analysis

Given the availability of numerous additional family members in FALSmq28, genome-wide parametric linkage analysis was performed. However, linkage analysis in this family using the currently available samples had insufficient power to identify any genomic region significantly linked to disease (LOD>3.3). This is attributable to the low availability of genotyping data from informative affected or obligate mutation carrying individuals (just three), and the high number of "at-risk" family members. These "at-risk" family members introduced a large degree of ambiguity. To combat

this, we employed liability classes based on age-dependent disease penetrance. These acted to inform the statistical model of the likelihood that an "at-risk" family member carrying the disease causal mutation would be affected by disease at their current age. However, these liability classes also carry a degree of uncertainty. As has been extensively described, highly variable age of disease onset and mutation penetrance levels are observed among ALS s, and further, the different ALS gene mutations have been observed to associate with varying ages of onset (as shown in Paper I, Chapter 4, Section 4.3.1). Together, these variances dictate that unique liability classes are likely to apply to each ALS mutation, and even each ALS family. As such, the averaged age-dependent penetrance liability classes employed here may not reflect the age-dependent penetrance of the ALS causal mutation in this family.

6.4.3 *In silico* pipeline for candidate mutation prioritisation

After having exhausted the genetic power of the four small families, long lists of candidate mutations remained. The causal mutation in these families may remain elusive until sufficient numbers of additional family members present with ALS, which may take decades. As such, alternate strategies are required to characterise genetically identified lists of candidate mutations, in order to prioritise which are most likely to cause disease. An *in silico* pipeline and associated scoring system was developed as part of this Chapter in order to achieve this goal in a consistent and unbiased manner. Additionally, our laboratory also has an *in vitro* pipeline in place to assess the functional characteristics of candidate mutations. The *in silico* pipeline developed here acts as a complementary tool to prioritise those candidate mutations most suited to *in vitro* analysis.

The *in silico* scoring system incorporated four characteristics including, gene expression, protein predictions, species conservation and genic tolerance. These characteristics were chosen owing to their correlations with known ALS gene mutations. In addition, the *in silico* tools used to assess each characteristic returned numeric values indicating a specific result which was not open to subjective interpretation by the end user. Further, each result could easily be converted to a numeric score to facilitate a straightforward scoring system, and subsequent rank.

Gene expression

Disease causal mutations must be expressed in tissue types affected by disease.

Therefore, we hypothesise that causal ALS mutations affect genes encoding proteins that are expressed in the brain and spinal cord. However, it remains possible that a gain-of-function mutation may cause a protein to be expressed in a different tissue type. Here, gene expression was assessed in the cerebellar cortex and spinal cord, which each contain motor neurons, and are affected in ALS patients. Both databases used here (HBT and GTex) are intended as reference resources and contain good quality expression data. However, the HBT provides data on age-related expression levels, while GTex provides a single age-averaged expression value. As such, expression was assessed at approximately 80 years of age (at which point the majority of ALS patients would have already presented with disease) in the cerebellar cortex, though expression in the spinal cord was an age-averaged value.

Protein prediction programs

Numerous protein prediction programs are available, each of which utilise a different algorithm and combination of gene and/or protein characteristics to predict the effect of a sequence variant. The characteristics assessed by these programs often include evolutionary conservation, location and context within the protein, and/or the biochemical consequence of the amino acid alteration (Richards et al., 2015). As such, each has its own strengths and weaknesses. Generally, these programs are 60-80% accurate for known pathogenic missense mutations (Thusberg et al., 2011), and most underperform for mutations with midler effects (Choi et al., 2012). To account for their differences and inaccuracies, it is considered prudent to utilise multiple prediction tools (MacArthur et al., 2014; Richards et al., 2015). Eight different programs were utilised during this project and multiple pathogenic predictions were required when prioritising the potential effect of a candidate variant.

Amino acid conservation across species

Protein residues that are conserved across species indicate that the amino acid is evolutionarily important and is likely to play an important role in protein structure, function and/or binding. Therefore, alterations to highly conserved amino acids are likely to have a detrimental effect on the protein. In turn, mutations affecting highly conserved residues are more likely to be pathogenic. This pathogenic effect may be from a toxic gain-of-function mechanism, or a loss-of-function mechanism that inhibits the effective functioning or binding of the protein.

Each of the three approaches used to assess amino acid conservation varied in complexity. The first was a simple manual approach that directly assessed whether

the substituted amino acid was shared by multiple species. Second, the PhastCons metric was calculated using a statistical model to identify conserved protein sequences by comparison of 18 different species including vertebrates, insects, bacteria and fungi (Siepel et al., 2005). Finally, the PhyloP metric utilised four statistical tests to assess both amino acid conservation and the rate of change across 36 mammalian species (Pollard et al., 2010). The conservation results of these approaches showed considerable variation. By incorporating the results of these different approaches, it was intended that the conservation score used in the *in silico* pipeline would provide a broad representation of the conservation of a candidate mutation affected residue across species.

Genic tolerance to variation

The natural variation of a gene is a measure of the frequency of neutral protein-altering sequence variants present in that gene. Genes that have high levels of natural variation (those containing many genetic variants) are said to have a higher tolerance for sequence changes without a negative effect on protein function. Conversely, genes that have low natural variation (few variants) are intolerant to variation, and therefore are constrained, indicating a crucial biological role, and low adaptability to variation.

Human genic tolerance is considered to be a better predictor of pathogenicity than conservation across species (Richards et al., 2015), protein prediction tools are imperfect (Thusberg et al., 2011), and gene expression may be altered by variation. Further, most known ALS genes have a low tolerance for variation. As such, genic tolerance was weighted more highly than any other characteristic as part of the *in silico* scoring system, being scored out of four, whereas the other characteristics were each scored out of two. Two different database scores were used to assess the genic tolerance of genes containing candidate mutations to avoid bias present in either database. The RVIS metrics consider all common functional variation, while the ExAC missense constraint score only accounts for missense variants. The overwhelming majority of known ALS causal mutations are non-synonymous/missense in nature, however other genetic variation has been reported to cause disease, including small indels as well as the pathogenic expansion in *C9orf72*. As such, while the tolerance of a gene for missense variants was most relevant when assessing non-synonymous candidate mutations, more generalised genic tolerance (still including tolerance to non-synonymous/missense variants) of a gene could not be discounted. Therefore, the use of these two databases should strike a good balance by primarily reflecting genic tolerance for missense mutations, and to a lesser extent that for other genetic variant

types.

Proof of principle

When applied to known ALS mutations and common benign variants, the *in silico* pipeline for assessment of potential pathogenicity showed a clear distinction in scores between the two categories (Table 6.5). This suggests that the pipeline can successfully distinguish between pathogenic ALS gene mutations and benign variation. Low scores of ∼2 were consistently generated for known benign variants and therefore setting the threshold for low priority variants at two was straight forward. Interestingly, we observed a greater variation between scores for the known ALS mutations. *SOD1* mutations scored considerably higher than less common ALS genes such as *CCNF* and *UBQLN2*. This could reflect why reduced penetrance is more commonly observed in families carrying these mutations, compared with highly penetrant *SOD1* mutations. Nonetheless, the scores of all ALS mutations were far greater than any benign variant. The threshold for high priority variants was set at five, as this was the closest round number to the *CCNF* score of ∼4.8, and exercised caution to not overstate the potential for pathogenicity. Those variants falling between these thresholds were classed as medium priority, as they exhibited some characteristics suggestive of pathogenic potential, but also some characteristics compatible with benign variation.

The proof-of-principle studies suggest that the scoring system developed here is a highly useful tool to aid in the selection of candidate mutations that warrant downstream *in vitro* or *in vivo* analysis for pathogenicity. Nevertheless, as more affected family members are recruited, or control samples screened, ongoing filtering may remove top ranked *in silico* candidates. As such, it is imperative that this scoring system is used as an adjunct tool to support genetic findings and guide downstream research, but cannot be used in place of additional genetic analysis as more family members are recruited. The re-identification of identical, or novel candidate mutations in the same gene in additional families and/or sporadic patients will also provide strong support for a causal role.

6.4.4 ALS families and their candidate mutations

6.4.4.1 FALS15

A total of 19 candidate mutations were identified in family FALS15. Five of these variants were classified as having a high potential for ALS pathogenicity following *in*

silico prioritisation. *CLCN4* p.I886T had the strongest support of these 19, with a score of eight (out of ten) using the *in silico* pipeline, and was also completely absent from all control databases. Interestingly, *CLCN4* is located on the X chromosome. As there is no male-to-male transmission evident in this pedigree, the pattern of inheritance of ALS in this family is compatible with the possibility of a dominant X-linked mutation. Indeed, the known ALS gene *UBQLN2* is a dominant X-linked gene (see Chapter 1, Section 1.4.1.5). As such, *CLCN4* may be another X-linked ALS gene. While the physiological role of this chloride channel gene remains largely unknown, it is likely to facilitate the transport of ions across intracellular membranes (Veeramah et al., 2013). The *in silico* assessment of potential pathogenicity applied here showed that *CLCN4* was highly intolerant to genetic variation and highly expressed in both brain and spinal cord. Additionly, mutations in *CLCN4* have been implicated as a cause of intellectual disability (Hu et al., 2016; Palmer et al., 2018), and have also been suggested as a potential cause of Epilepsy (Veeramah et al., 2013). While these conditions are not neurodegenerative, they do affect neuronal tissue, suggesting that alteration of *CLCN4* has a detrimental effect on this ALS relevant tissue type. Combined, this supports the potential for *CLCN4* p.I886T to cause ALS in this family.

The other four high priority candidates in FALS15 were all autosomal, heterozygous variants in both the proband and his obligate mutation carrier mother. Each resides on a separate chromosome. Of these four variants, the *SCN4A* variant may be a rare variant present in the general population. While the *SCN4A* gene showed a moderate level of intolerance to variation, this particular variant was present in six control individuals in gnomAD, including four NFE individuals, though it was filtered from the database as a low quality variant call. However, in the WES data from the two FALS15 family members, this variant had a high quality score (GQ=99) and was validated by Sanger sequencing. Without validation of the variant calls in the gnomAD controls, it is impossible to confirm whether this variant is actually a rare population-based variant. Interestingly, the fifth ranked candidate mutation *SUPV3L1* p.Q168E, is an attractive candidate as this gene encodes a DNA- and RNA- binding protein (like several other ALS genes) that is known to interact with HNRNPA1, a known ALS protein.

6.4.4.2 FALS45

Eleven candidate mutations were identified in family FALS45, of which six were determined by the *in silico* pipeline to have a high potential for ALS pathogenicity. The top ranked candidate was *SCCPDH* p.V256L with a score of 6.7 (out of 10).

Little is known about the function of the protein encoded by *SCCPDH*. However, given that *SCCPDH* was highly expressed in the brain and spinal cord, it is likely to have a role in the nervous system. As such, *in vitro* and *in vivo* analyses will be necessary to elucidate its potential effect on neuronal functions, and its potential contribution to ALS.

The high priority candidate, *GRIN2D* p.V144L, was absent from all databases and was shown to be highly intolerant to variation. Its encoded protein is a subunit of the N-methyl-D-aspartate (NMDA) receptor, an ionotropic glutamate receptor. NMDA receptors facilitate synaptic transmission and have been shown to have crucial roles in brain development, memory formation, synaptic plasticity and neurotoxicity (Laube et al., 1997; Nakanishi, 1992; Olney, 1990). NMDA receptors, and *GRIN2D* specifically, have previously been linked to neurodegenerative disease. In Alzheimer's disease, NDMA receptor regulation and activation has been implicated in disease-related synaptic dysfunction (reviewed in Mota et al., 2014). In Parkinson's disease, NMDA receptors have been found to be more abundant in the striatum of patients compared to controls (Weihmuller et al., 1992), while *GRIN2D* expression is increased in peripheral blood samples from patients compared with controls (Liu et al., 2016). Taken together, *GRIN2D* p.V144L is a strong candidate mutation in FALS45.

Four of the six high-priority candidate mutations (including the top ranked *SCCPDH* p.V256L) were also present in either one or two individuals from a control database. Three of these four (again, including *SCCPDH* p.V256L) were also found in another ALS patient, in addition to the control individuals. While it is possible that the controls harbouring any of these candidate mutations may go on to develop ALS, or that the variant call may be a sequencing artefact in the control databases (due to lack of validation), it is probable that these are actually rare variants in the population. This is further supported by the fact that each of the relevant genes has an average level of genic tolerance and therefore has the potential to adapt to this variation without adverse consequences.

Of all eleven candidate mutations, the bottom ranked *ZNF132* had the most ALS-relevant known gene function, as a nucleic acid binding protein. However, as this candidate mutation is present in five gnomAD/ExAC control individuals (albeit that only one is from the NFE population), coupled with a lack of any other supportive evidence, it is unlikely to cause ALS in this family.

6.4.4.3 FALSmq2

Among the 16 candidate mutations identified in family FALSmq2, just one, *STRN4* p.D362E, was assessed to have a high potential for ALS pathogenicity. However, this variant was present in one control individual in the Project MinE database. Little is known about the encoded protein, STRN4, other than that it binds calmodulin. Calmodulin is a ubiquitous and highly abundant protein with hundreds of protein targets and is involved in numerous cellular functions. In the absence of additional genetic data, *in vitro* studies would be required to shed light on the potential involvement of STRN4 in the cause of ALS.

Interestingly, the *ZFHX2* p.T565Rfs*19 candidate mutation, ranked third by *in silico* scoring, is a zinc finger homeobox protein involved in nucleic acid binding. Initially, this variant was reported as a single nucleotide variant by the bioinformatic pipeline but was found to cause a frameshift upon direct validation. The ZFHX2 protein is 2,572 amino acids long and this frameshift was predicted to cause nonsense mediated decay of the mRNA, using MutationTaster2 (Schwarz et al., 2014). This may lead to haploinsufficiency, and possibly a loss-of-function for the ZFHX2 protein. Other than MutationTaster2, all other protein prediction programs utilised here were limited to missense substitutions or stop-gain mutations, and therefore were unable to score this frameshift candidate mutation. The nonsense mediated decay prediction by MutationTaster2 together with the lack of compatibility with the other prediction programs led to an assignment of a full score of two points for the protein prediction criteria for this candidate mutation. Unfortunately, no data was available for genic tolerance and therefore a score of zero was assigned for this characteristic. Additionally, the amino acid conservation score was applied to the single residue at which the frameshift occurred, however as over 80% of the protein was affected, this may not be a true representation of the lack of conservation introduced by this candidate. As such, an incomplete *in silico* assessment of pathogenicity was completed for this candidate mutation, which may have artificially reduced its score and associated ranking.

6.4.4.4 FALSmq20

Sixty-four candidate mutations were identified for family FALSmq20. *In silico* assessment showed that eleven candidate mutations had a high potential for ALS pathogenicity. The top ranked candidate, *RASGRF1* p.S34W, scored very highly at

9.4 (out of 10) and was completely absent from all screened control cohorts. The gene is extremely intolerant to variation, highly expressed in neuronal tissue and relatively well conserved. The encoded protein is a Guanine nucleotide exchange factor, which activates the RAS protein and is primarily expressed in adult neurons. It is involved in regulating cellular processes such as cell proliferation and differentiation. While *RASGRF1* is a known disease gene in myopia, it has also been linked to neurodegeneration. A *RASGRF1* knockout mouse model showed significant differential expression of genes related to neurodegenerative processes affecting memory and learning pathways (Fernandez-Medarde et al., 2007). Further, a *RASGRF1* genetic variant has also been associated with increased memory performance in humans (Barman et al., 2014). While these are not motor deficits, these links with memory formation may indicate a role in neurodegeneration. Taken together, this supports further assessment of *RASGRF1* in this family.

6.4.4.5 FALSmq28

While no candidate mutations in coding sequence were identified in either WES or WGS data from family FALSmq28 (Analysis 1), it was possible to exclude ~86% of the genome as being linked to ALS in this family using genome-wide linkage analysis. A total of 41 genomic regions totalling ~73 Mb remain as potentially harbouring the ALS causal mutation. However, the highest LOD score was just 1.1924, therefore no genomic region was significantly linked to disease in FALSmq28. Family-based analysis of genomic regions that were not excluded by linkage analysis failed to identify any candidate mutations. Nevertheless, by reducing the search for the ALS causal mutation in FALSmq28 to just ~14% of the genome, the scope of genetic analysis in this family has been substantially reduced by these efforts.

It must be noted that the genetic linkage analysis model used here assumed an autosomal dominant inheritance pattern of disease. However, recessive inheritance cannot be excluded, nor can the possibility that the patients are two sporadic cases. If we consider this familial disease, autosomal inheritance is evident because there was male-to-male transmission of the disease allele. While it is not clear that dominant inheritance is at play in this family, the vast majority of ALS gene mutations show autosomal dominant inheritance, with recessive mutations rarely observed. Given the low prevalence of ALS, recessive ALS mutations are generally only seen in consanguineous families. No evidence of consanguinity was apparent in FALSmq28, particularly between the parents of the ALS patients. The inheritance pattern is

compatible with an autosomal dominant mutation with incomplete disease penetrance, which is a common feature of ALS families. As such, it was deemed reasonable to apply an autosomal dominant inheritance model as part of genetic linkage analysis.

Interestingly, over 99% of the variants identified by WES in FALSmq28 were shared by all three family members, while this figure was just 35-40% using WGS data. It is likely that this result reflects the increased conservation of coding regions compared with non-coding regions, particularly intergenic regions that account for ~56% of WGS variants in this family. Additionally, the increased sequencing coverage of WES (100X) compared with WGS (30X) led to higher confidence variant calls in the WES data set. The increased proportion of false positive variant calls in the WGS dataset would therefore have reduced the proportion of shared variants.

Alarmingly, Sanger sequencing validation of WES- or WGS-derived candidate mutations in this family showed 16 of 16, and 13 of 14 to be false positive variant identifications. Notably, one of these false positive variants was identified by both WES and WGS. All false positive variants were identified within genomic regions that were highly repetitive and/or duplicated. Chapter 8, Section 8.3.3 will discuss the issue of NGS false positive variant calls in detail. The candidate mutations identified by the two sequencing technologies (that were subsequently found to be sequencing artefacts) showed minimal overlap, with just a single regulatory variant from Analysis 2 being called from both WES and WGS data, though it too was later found to be a sequencing artefact. This suggests that unique factors cause different artefacts between the two technologies. These factors are likely related to the library preparation/capture phase, as the sequencing and bioinformatics pipelines applied to each of the WES and WGS datasets were the same. As such, the false positive candidate mutation identified by both WES and WGS is likely to be an artefact of the sequencing chemistry, or the bioinformatics processing algorithms. In Chapter 8, Section 8.3, in particular Section 8.3.1, the advantages and disadvantages of WES and WGS will be discussed in detail.

Sequencing validation also revealed that the single WGS-derived candidate mutation that withstood bioinformatic filtering in family FALSmq28, a homozygous variant upstream of *MIR512*, was actually a population-based variant. Sanger sequencing of seven unrelated Australian controls showed three had a homozygous genotype identical to that seen in the affected and obligate mutation carriers from FALSmq28, while two more control individuals carried the variant in a heterozygote state. This

variant was also present as a high quality, heterozygous variant in WGS data from 38 individuals of African descent in the gnomAD control database. It was however absent from all WES data from control databases, as it falls well outside of the exome. Additionally, 497 heterozygous and 293 homozygous individuals were identified with high quality WGS genotypes among 850 Australian ALS/FTD patients in the 850-sample VCF (described in Chapter 2, Section 2.1.3). Together, these findings indicate this is a common population-based variant. Indeed, it has also been added to the most recent release of dbSNP (dbSNP150) that was not available during the analysis phase of this candidature. Further, it is possible that the frequency of this variant is under-represented in WGS control databases. Both gnomAD and MGRB (from which this variant is absent) report a 2bp insertion (including the alternate A allele) at this same position, as a heterozygous and homozygous variant with a MAF>0.4. Given that no validation data is available from either database, and the innate differences between NGS variant calling tools, it is possible that some of these variant calls are incorrect (to be discussed further in Chapter 8, Section 8.3 , particularly Section 8.3.2), and actually represent the single base variant reported here.

We have exhausted all avenues to identify nucleotide level candidate mutations in this family with the existing datasets. If such a variant is causing ALS in FALSmq28, the only possibilities are that the causal mutation was masked by a sequencing artefact of WES and/or WGS, is a low quality variant that was filtered in Analysis 3, or that it has been reported in a control database in three or more individuals (see Section 6.4.2.2 above for an explanation of why such a mutation would be in a control database). If a sequencing artefact has masked the causal ALS mutation from analysis in FALSmq28, this may be explained by the mutation falling in a region not captured or covered by WES or WGS, inadequate coverage or an incorrect variant call caused by bioinformatics processing. Chapter 8, Section 8.3.3 will provide an indepth discussion of the possible sources of NGS sequencing artefacts. Alternatively, ALS is being caused by a different mutation type in this family, such as a structural variant (SV) (such as a copy number variant (CNV)), that has not been captured by WGS or WES. In Chapter 8, Section 8.5 we will discuss the planned investigation of CNVs and SVs as a cause of ALS, including in FALSmq28.

"'The thing about growing up with Fred and George,' said Ginny thoughtfully, 'is that you sort of start thinking anything's possible if you've got enough nerve.'"

JK Rowling - Harry Potter and the Half Blood Prince

7

Searching for genetic differences between ALS-discordant monozygotic twins

7.1 Introduction

This Chapter addresses the second part of Aim 3 of this thesis; to identify novel ALS genes and mutations in monozygotic twins discordant for disease. Monozygotic twins that are discordant for ALS offer a rare opportunity to identify potential genetic, epigenetic or environmental factors that underlie disease discordance. ALS-discordant monozygotic twin/triplet sets (both familial and sporadic) were screened for *de novo* mutations that may underlie the onset or variable penetrance of ALS. As part of this project, DNA samples were available from three monozygotic twin pairs, and one monozygotic triplet set, each of which consisted of one ALS patient and their unaffected co-twin/triplets. Two twin pairs consisted of one SALS patient and their unaffected co-twin. The triplet set and the other twin pair were from families with a history of ALS, carrying a *SOD1* mutation and *C9orf72* expansion, respectively. While all three triplets and both twins carried their respective family mutations, just one of each set had developed ALS at the time of analysis. This cohort of ALS-discordant monozygotic twins represented a unique resource for uncovering novel genetic factors contributing to ALS pathogenesis.

Monozygotic (MZ) twins result from a single fertilisation event where one zygote has split into two embryos, so that both twins have an identical genetic code. On the other hand, dizygotic (DZ) twins develop from two separate ova, each of which has been fertilised by a distinct sperm cell, and are thus like any other pair of siblings, sharing an average 50% of their DNA sequence. Beyond these genetic characteristics, both MZ and DZ twins share their age, pre-natal environment, and in most cases, where twins have been raised together, partially share their post-natal environment. Owing to these characteristics, twins have long been utilised in heritability studies to estimate the contribution of genetics to a phenotypic trait. That is, the phenotypic differences between MZ twins should be attributable to distinct environmental factors, whereas those between DZ twins may be caused by either genetic or environmental factors, or a combination of both (Boomsma, 2013). As such, the extent of phenotypic similarity between MZ twins compared to that of DZ twins reflects the degree of genetic influence over a trait.

As technology has advanced, the utility of twin studies in other areas of research has also become apparent. This is particularly true for disease-discordant MZ twin pairs, in which one co-twin is affected by disease, while the other remains unaffected. Such twin pairs have emerged as a unique resource to identify molecular factors contributing to the cause of disease in the absence of confounding genetic variation (Zwijnenburg et al., 2010). This approach has been utilised across the "omics" research space, including as part of genomic, epigenomic, transcriptomic, proteomic and metabolomic studies (van Dongen et al., 2012).

While MZ twins are considered to be genetically identical, there exists the possibility that *de novo* mutations may distinguish one co-twin from the other. Indeed, early post-zygotic *de novo* mutations have been found to substantially contribute to the aggregate of all *de novo* mutations present within an individual, at a rate of $0.04\text{-}0.34\times10^{-8}$ (Dal et al., 2014). As such, early post-zygotic *de novo* mutations may underlie disease discordance in MZ twins. Disease-discordant twins have undergone comparisons using NGS data to identify such disease causal *de novo* mutations for Van der Woude syndrome (Kondo et al., 2002), Schizophrenia (Castellani et al., 2017; Reble et al., 2017), Neurofibromatosis type 1 (Vogt et al., 2011) and Frontotemporal dysplasia (Robertson et al., 2006). In Van der Woude syndrome, Kondo et al. (2002) identified a *de novo IRF6* mutation in a disease-discordant MZ twin pair and went on to identify *IRF6* mutations in 45 families with the same condition and in 13

additional families with the closely related condition, Popliteal pterygium syndrome, to confirm that *IRF6* mutations are a major cause of these syndromes. Thus, there exists exciting potential for disease-discordant twin studies to identify novel, and widely applicable, causes of disease.

Given the current difficulties in identifying the remaining ALS genes, gene discovery approaches using ALS-discordant MZ twins offer an alternative approach to identify novel causes of ALS. Further, *de novo* mutations have previously been implicated in ALS. Chesi et al. (2013) conducted a screen of 47 SALS trios (SALS affected patients and their unaffected parents) to identify 25 novel missense *de novo* mutations in patients, and subsequently implicate *SS18L1/CREST* as a novel ALS gene. This suggests that other *de novo* mutations may also be contributing to the cause of ALS, and that their identification may lead to a better appreciation of the genetic spectrum of disease. The high heritability estimates for all forms of ALS (Al-Chalabi et al., 2010; McLaughlin et al., 2015) also suggest it is possible that post-zygotic *de novo* mutations between MZ twins may be a cause of disease discordance. *De novo* mutations may also modify the onset or phenotypic presentation of ALS between co-twins and thereby implicate genes or other loci that contribute to phenotypic variability among ALS patients (as described in Chapter 1, Sections 1.3.1 and 1.6.1). Indeed, such phenotypic modifier variants have been identified for conditions such as Duchenne muscular dystrophy (Bello et al., 2016) and Huntington's disease (Beanovi et al., 2015).

The manuscripts presented in this Chapter utilised the ALS-discordant MZ twin approach to search for novel genetic causes of ALS. This includes a first-author manuscript that describes whole-genome sequencing (WGS) analysis together with extensive validation and bioinformatics strategies, to search for nucleotide level *de novo* mutations between co-twins/triplets that may underlie disease discordance, and represent novel ALS genes. Additionally, a co-authored manuscript describes an investigation of the epigenetic and transcriptomic profiles of these ALS-discordant twin sets.

7.2 Manuscripts

7.2.1 Manuscript III – Identifying *de novo* variants between ALS-discordant monozygotic twins

The study presented in Manuscript III sought to utilise the disease-discordant MZ twin model to identify novel genetic causes or modifiers of ALS. It was hypothesised that the affected individual in each twin pair discordant for SALS may harbour a *de novo* mutation that caused disease. On the other hand, it was hypothesised that the affected individual in each twin/triplet set that was discordant for FALS may carry genetic variants that modify the phenotypic manifestation of disease (i.e. early or late onset) given that both the affected and unaffected co-twins/triplets carried known ALS causal mutations.

In order to identify such genetic contributors to disease, WGS was performed for all co-twins/triplets. Analysis focused on nucleotide level variation, given that all but two of the hundreds of known ALS gene mutations are either SNP or indel variants. Rather than solely focusing on the exome, both coding and non-coding regions were considered, as disease modifying variants may affect important non-coding regulatory regions that influence gene expression.

The code in Appendix A.2.20 was applied to WGS data from each twin/triplet pair to identify discordant variants (high confidence variants with a genotype that differed between co-twins/triplets). This analysis identified tens of thousands of discordant variants between each pair of affected and unaffected co-twins/triplets. This was startling, given that all twin/triplet sets were previously shown to be MZ using SNP microarrays and microsatellite genotyping (performed by the candidate during Master of Research candidature). For the triplet set, the affected triplet was separately compared to each unaffected triplet.

Three independent validation strategies were then applied to determine whether the putative discordant variants were truly present between co-twins/triplets. First, 24 putative discordant variants were randomly selected to undergo direct DNA sequencing for validation (as per Chapter 2, Section 2.4.2). This found that all 24 putative discordant variants were actually concordant between co-twins/triplets, suggesting that all were not truly discordant. The second validation approach utilised SNP microarray genotype data (generated as per Chapter 2, Section 2.3). This

analysis involved 1) combining all the putative discordant variants identified for each twin/triplet pair, 2) identifying any database SNPs (rsID variants) among the putative discordant variants, 3) identifying which of these had been genotyped using the microarray and 4) extracting and comparing genotype data for each twin/triplet pair for each of these variants. The Custom Script in Appendix A.2.21 and the script in Appendix A.2.22 were developed and applied for this purpose. This analysis found that of the 81 putative discordant variants (across all twin/triplet pairs) for which SNP microarray genotype data was available, all had concordant SNP microarray genotypes between co-twins/triplets, again suggesting that all were not truly discordant. Lastly, re-sequencing of the genome was performed for one twin set as a replicate analysis. Discordant variants were identified from this new WGS data set, again using the code in Appendix A.2.20. The Custom Script 3.6 was then applied to the putative discordant variants identified from each of the two WGS datasets for this twin set, to identify any shared discordant variants. While 18,599 and 3,543 putative discordant variants were identified in the original and re-sequenced WGS datasets, none were common to both datasets. This suggested that all putative discordant variants identified using WGS in this twin pair, were in fact artefacts of the WGS process rather than true discordant variants. Due to cost restraints, re-sequencing was not possible for the other twin/triplet sets.

Given the high false discovery rate of putative discordant variants from WGS, bioinformatics processing, namely alignment and variant calling, of the original raw WGS data (from all four twin/triplet pairs), was repeated using three additional processing pipelines in an attempt to identify any true *de novo* mutations between co-twins/triplets. These pipelines were applied by two separate service providers, and employed different versions of the Burrows Wheeler Alignment (Li and Durbin, 2009, 2010) and Genome Analysis ToolKit (McKenna et al., 2010) (BWA-GATK) tools originally used for raw WGS data processing, and two different versions of the Isaac alignment and variant calling software (Raczy et al., 2013). For each twin/triplet pair, each of the four processed datasets was analysed for discordant variants, again using the code in Appendix A.2.20. This resulted in four lists of putative discordant variants for each twin/triplet pair, which were then intersected using the Custom Script 3.6, to identify variants that overlapped between the pipelines. The Custom Script in Appendix A.2.24 was applied to generate Venn diagrams and identify any overlap of the four putative discordant variant sets for each twin/triplet pair. No putative discordant variants in any twin/triplet pair were shared by all four pipelines. However, three of the four pipelines did show a small overlap of between 0.03-3.2% of

putative discordant variants for each twin set.

To evaluate the likelihood that these overlapping variants were truly discordant, the Custom Script 3.7 was applied to determine whether overlapping variants were from confidently "callable" genomic regions. Confidently "callable" regions were previously defined according to extensive replication and comparison of WGS data for a single individual across five sequencing platforms, seven alignment tools and three variant calling tools (Zook et al., 2014). The analysis here showed that all putative discordant variants overlapping between processing pipelines fell outside of the confidently "callable" genome, and were unreliable variant identifications. It was thus concluded that WGS had not detected any post-zygotic, nucleotide level *de novo* mutations that caused or modified the presentation of ALS in these four twin/triplet sets.

The Custom Script in Appendix A.2.23 was used to determine the distribution of putative discordant variants between SNP and indel variant types. This showed that while the number of SNP and indel putative discordant variants identified by the BWA-GATK pipelines were proportionally as expected given their abundance in the genome (∼79% and ∼21%, Mullaney et al., 2010), indel variants were over-represented among the putative discordant variants identified by the Isaac pipelines, accounting for more than 90%. This suggests that Isaac processing may be less reliable for calling indel variants than GATK. Further, the distribution of putative discordant variants across the functional classes of the genome was determined using a variation of the Custom Script in Appendix A.2.19. Unsurprisingly, this showed that ∼80% of putative discordant variants were intergenic or intronic variants, and less than 1% fell within coding regions.

Author contributions

The candidate designed all analyses, performed all bioinformatics analyses of discordant variants, conducted Sanger sequencing validation experiments, designed all primer sequences, completed all statistical analyses and wrote the manuscript. NT modified and ran bioinformatics scripts used to identify discordant variants and contributed to study design. DB wrote the original bioinformatics scripts used to identify discordant variants. NG performed Sanger sequencing validation experiments. KW conceptualised the project, contributed to study design and performed SNP microarray validation. IB provided intellectual input and supervised the study. All authors contributed to the editing of the manuscript.

Whole genome sequencing of amyotrophic lateral sclerosis discordant monozygotic twins identifies thousands of false positive *de novo* mutations

Emily P. McCann[1], Natalie A. Twine[1,2], Dennis C. Bauer[2], Natalie Grima[1], Garth A. Nicholson[1,3,4,5], Dominic B. Rowe[1], Ian P. Blair[1], Kelly L. Williams[1]*

[1] Macquarie University Centre for Motor Neuron Disease Research, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia

[2] Transformational Bioinformatics, Commonwealth Scientific and Industrial Research Organisation, Sydney, New South Wales, Australia.

[3] Northcott Neuroscience Laboratory, ANZAC Research Institute, Sydney, New South Wales, Australia

[4] Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia

[5] Molecular Medicine Laboratory, Concord Hospital, Concord, New South Wales, Australia

*Corresponding author:

Email: kelly.williams@mq.edu.au (KW)

ORCID: 0000-0001-6319-9473

Email addresses:

Emily P. McCann: emily.mccann@mq.edu.au

Natalie A. Twine: natalie.twine@csiro.au

Dennis C. Bauer: denis.bauer@csiro.au

Natalie Grima: natalie.grima@mq.edu.au

Garth A. Nicholson: garth.nicholson@sydney.edu.au

Dominic B. Rowe: dominicrowe@mac.com

Ian P. Blair: ian.blair@mq.edu.au

Kelly L. Williams: kelly.williams@mq.edu.au

Abstract

Background

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease, that causes progressive muscle weakness, wasting and spasticity, leading to the loss of the ability to walk, speak and eventually, breathe. To-date, the only proven causes of ALS are gene mutations. Considerable phenotypic variation is evident between ALS patients, including those that carry identical causal gene mutations. Disease discordant monozygotic twins provide a unique opportunity to study phenotypic variation. Somatic *de novo* variants may exist between discordant co-twins that act as causal ALS mutations or phenotypic modifiers. Whole genome sequencing (WGS) was performed in three Australian monozygotic twin sets and one monozygotic triplet set, all discordant for ALS, in order to identify discordant variants that represent somatic *de novo* mutations between co-twins/triplets. One monozygotic triplet set carried a pathogenic *SOD1* p.I114T mutation, and one monozygotic twin set harboured a pathogenic *C9orf72* hexanucleotide repeat expansion.

Results

Initial WGS analysis suggested that tens of thousands of discordant variants existed between co-twins/triplets, but failure to validate selected variants indicated that these were artefacts of WGS. To successfully identify bona fide discordant variants within a twin set, four independent bioinformatic data processing pipelines were applied to the raw sequence read data to remove false discordant variants. Intersection of the discordant variants from each of the four processed datasets showed that >98% of putative discordant variants were only present in one dataset and were therefore artefacts of

bioinformatics processing. The remaining <2% of putative discordant variants were present in genomic regions that are notoriously enriched for sequencing artefacts, and were thus uninformative.

Conclusions

No bona fide somatic *de novo* mutations were identified in peripheral blood-derived WGS data from any of the four Australian ALS discordant MZ twin/triplet sets. Striking discrepancies were observed between the different bioinformatics processing pipelines that were applied to the WGS data, which highlights the importance of independent validation of variants identified by WGS.

Keywords

Amyotrophic lateral sclerosis, monozygotic twins, whole genome sequencing, disease discordance, false positive

Background

Amyotrophic lateral sclerosis (ALS; also known as motor neuron disease, MND), is a fatal late onset, rapidly progressive neurodegenerative disease characterised by degeneration of both the upper and lower motor neurons. Patients experience progressive muscle weakness, wasting and spasticity leading to loss of the ability to walk, speak, eat and eventually breathe. Most patients die from respiratory failure within just two to five years from symptom onset. Ten percent of ALS patients have a family history of disease (familial ALS; FALS), while the remaining 90% of cases occur seemingly sporadically (sporadic ALS; SALS). To date, there is no effective treatment for ALS, and little is known about disease pathogenesis. Gene mutations are the only proven cause of ALS. Missense mutations in *SOD1* and a pathogenic repeat expansion in *C9orf72* are the two most common genetic causes of Australian ALS accounting for more than 50% of FALS and <5% of SALS [1, 2].

Genetic and phenotypic heterogeneity is observed amongst ALS patients. A highly variable disease course is apparent among patients in terms of age and site of onset, and disease progression. The onset of classical ALS can range from the second to ninth decade of life [3], while the site of onset may be in any limb, the bulbar musculature, and in rare cases the trunk [4]. Comorbidity with frontotemporal dementia (FTD) is observed in ~20% of ALS patients [5-7]. Further, ALS may progress rapidly or slowly, so that disease duration may be 3 months to over 10 years [3]. Phenotypic variation is also apparent between patients with identical causal gene mutations, including those within the same ALS family. The rate of ALS discordance may be as high as 90% in

126 monozygotic (MZ) twins [8-11]. This suggests that modifying factors may be
127 contributing to the variable ALS phenotype.
128
129 MZ twins facilitate powerful study design in genetic research. Since MZ twins
130 arise from a single zygote they theoretically share an identical set of genetic
131 information in all cells. Studies of twin pairs, whether MZ or dizygotic (DZ;
132 arising from two separate zygotes), can avoid confounding factors such as
133 age, early development and environmental exposure. The phenotypic
134 concordance between MZ and DZ twin sets has been utilised extensively in
135 heritability studies to estimate the extent of the genetic contribution to a trait
136 [12]. In ALS, heritability studies in twins have estimated that 60% of sporadic
137 disease risk is attributable to genetic factors [10]. The utility of twin studies
138 has grown with advances in next-generation sequencing (NGS) technologies
139 [13]. In particular, the identification of molecular differences between disease
140 discordant MZ twins has emerged as an exciting avenue for the discovery of
141 novel disease causing or modifying factors [14].
142
143 While MZ twins are assumed to be genetically identical, early post-zygotic
144 mutations have been found to contribute a substantial proportion of *de novo*
145 mutations found within an individual, albeit at a low rate of $0.04\text{-}0.34 \times 10^{-8}$,
146 representing between one and ten *de novo* mutations per individual [15]. Such
147 early post-zygotic *de novo* mutations may underlie disease discordance
148 between MZ twins. For example, single nucleotide *de novo* mutations have
149 been implicated in MZ twins discordant for Van der Woude syndrome [16],

Schizophrenia [17, 18], Neurofibromatosis type 1 [19] and Frontotemporal dysplasia [20].

Each ALS discordant MZ twin pair provides a unique opportunity to discover a potential novel molecular cause or modifier of disease. *De novo* mutations in ALS discordant MZ twins may be a cause of disease or affect the presentation of clinical characteristics. In the current study, we performed whole genome sequencing (WGS) of three MZ twin sets and one MZ triplet set discordant for ALS, to seek discordant variants that represent *de novo* somatic mutations contributing to the aetiology of ALS. While no disease-associated *de novo* mutations were identified, the analyses revealed a startling number of false positive discordant variants likely sequencing artefacts, in the WGS data. This prompted an extended analysis that involved the comparison of discordant variants that were identified using four simultaneous but independent bioinformatics processing pipelines for the WGS data. This supported the absence of informative discordant variants between co-twins/triplets, but also highlighted the abundance of sequence artefacts introduced to WGS datasets by bioinformatics processing pipelines.

Results

Monozygotic twins and triplets discordant for ALS

Four Australian twin/triplet sets were discordant for ALS, whereby one twin/triplet had ALS and their co-twin/triplets were unaffected by disease (pedigrees provided in Fig. 1). Zygosity testing using existing SNP (single nucleotide polymorphism) microarrays confirmed all twin/triplet sets were

monozygotic. Two sets had a family history of ALS. Each of the triplet set harboured a *SOD1* p.I114T mutation, while each of a twin set harboured a pathogenic *C9orf72* hexanucleotide repeat expansion. Clinical details of each twin/triplet set are provided in Table 1.

Table 1. Clinical details of the ALS discordant twin/triplet sets.

MZ set	ALS	Status	Sex	Mutation	Age of onset	Duration (months)
Female *SOD1* triplet set	FALS	ALS	F	*SOD1* p.I114T	50	Unknown
		Asymptomatic	F	*SOD1* p.I114T		
		Asymptomatic	F	*SOD1* p.I114T		
Male *C9orf72* twin set	FALS	ALS	M	*C9orf72 HRE*	52	36
		Asymptomatic	M	*C9orf72 HRE*		
Female SALS twin set	SALS	ALS	F		42.7	Alive at 51 months
		Unaffected	F			
Male SALS twin set	SALS	ALS	M		78.5	28.4
		Unaffected	M			

HRE: hexanucleotide repeat expansion

Whole genome sequencing of ALS discordant twins identified thousands of false positive *de novo* (discordant) mutations

All four twin/triplet sets underwent WGS. One twin set underwent re-sequencing of the identical DNA sample at a second sequencing provider as

186 a validation step. Sequencing quality metrics are provided in Table 2 and are
187 similar across all samples.
188
189 Table 2. Whole genome sequencing raw data quality metrics.

MZ set	Status	Sequencing provider^	Sequencing prep	Sequencing yield (bases)	Throughput mean depth
Female *SOD1* triplet set	ALS	KCCG	Illumina PCR-free	113,635	39.8
	Asymptomatic	KCCG	Illumina PCR-free	124,116	43.4
	Asymptomatic	KCCG	Illumina PCR-free	144,546	41.9
Male *C9orf72* twin set	ALS	KCCG	Illumina Nano	142,824	50.0
	Asymptomatic	KCCG	Illumina Nano	146,824	51.4
Female SALS twin set*	ALS	KCCG	Illumina PCR-free	140,312	49.1
	Unaffected	KCCG	Illumina PCR-free	144,896	50.7
Male SALS twin set	ALS	KCCG	Illumina PCR-free	145,652	41.7
	Unaffected	KCCG	Illumina PCR-free	152,581	41.5
Female SALS twin set*	ALS	Macrogen	Illumina PCR-free	147,209	51.5
	Unaffected	Macrogen	Illumina PCR-free	149,775	52.4

*Identical DNA samples were sequenced twice at two different sequencing providers
^KCCG, Kinghorn Centre for Clinical Genomics (Sydney, Australia); Macrogen (Seoul, Korea)

190
191 Discordant variants were defined as genomic sites with a called genotype and
192 a coverage score greater than 30, at which the genotype call differed between
193 the ALS affected co-twin/triplet and their unaffected co-twin/triplet. Variant call
194 files (VCFs) processed using the genome analysis toolkit (GATK) and the

associated best practices [21] (according to pipeline 1, Table 3) were utilised to identify discordant variants. A two-sample VCF for each twin/triplet pair was subsetted from a large 850-sample joint-called VCF, containing data from the 11 twin/triplets under analysis as well as an additional 839 Australian ALS and FTD patients. Custom python scripts were then applied to identify discordant variants from the two-sample VCF for each twin/triplet pair. Importantly, when considering the *SOD1* triplet set, discordant variants were identified for each of the two possible pairings of the affected triplet with an unaffected triplet. That is, triplet analysis A compared the affected triplet with one unaffected triplet, and triplet analysis B compared the affected triplet to the alternate unaffected triplet.

Table 3. Details of the bioinformatics processing pipelines applied to raw WGS data.

Process	Sequencing provider 1				Sequencing provider 2			
	Pipeline 1		Pipeline 2		Pipeline 3		Pipeline 4	
	Software	Version	Software	Version	Software	Version	Software	Version
Aligner	BWA mem	v0.7.15	Isaac Aligner	00776.15.01.27	BWA mem	v0.7.10	Isaac Aligner	v01.15.02.08
Variant Caller	GATK Haplotype Caller	v3.7	Isaac Variant Caller	starka-2.1.4.2	GATK Haplotype Caller	v3.4	Isaac Variant Caller	v2.0.13
Merge per sample gVCFs	GATK Combine GVCFs	v3.7	GATK Combine GVCFs	v3.7	N/A	N/A	N/A	N/A
Genotype all samples	GATK Genotype GVCFs	v3.7	GATK Genotype GVCFs	v3.7	N/A	N/A	N/A	N/A

Comparison of WGS data between the ALS affected co-twin/triplet and their unaffected co-twin/triplet identified the following number of discordant variants

within a twin/triplet set: 12,240 (*SOD1* triplet pair A), 14,097 (*SOD1* triplet pair B), 55,132 (*C9orf72* twins), 18,599 (female SALS twins), and 30,994 (male SALS twins). Over 87% of these discordant variants were identified in intergenic or intronic regions, with less than 1% were found in exonic regions. The distribution of discordant variants between the various genomic functional classes (exonic, intergenic, intronic, non-coding RNA, (ncRNA), splicing, upstream/downstream and untranslated region (UTR)) is depicted in Figure S1 (see Additional file 1).

Failure to validate discordant variants suggested the majority are sequencing artefacts

As the number of discordant variants identified from WGS data was significantly greater than expected, given the known *de novo* mutation rate (typically one to ten per individual), we sought to validate a subset of discordant variants across all five twin/triplet pairs using existing SNP microarrays and direct Sanger sequencing. For one twin set, this was extended to re-sequencing of the genome using the same DNA samples used in the first round of WGS.

Eighty-one putative discordant variants identified across the five twin/triplet pairs had been previously genotyped using SNP microarrays. Genotype assessment of these variants showed that all had concordant SNP microarray genotypes between co-twins/triplets, strongly suggesting that all 81 putative discordant variants were false (i.e. not discordant). Similarly, Sanger sequencing of 24 selected putative discordant variants across the twin/triplet

sets also showed genotype concordance between co-twins/triplets. Re-sequencing of the genome of the female SALS twin set, at a different sequencing provider but using the same library preparation method, sequencing instrument, and variant identification pipeline as used in the first round of WGS, identified 3,543 discordant variants. However, there was no overlap between this set of discordant variants and that identified in the first round of WGS. Therefore, re-sequencing of the genome failed to validate any of the original 18,599 discordant variants in this twin pair.

Three independent validation methods failed to confirm any *de novo* mutation. It is therefore likely that all putative discordant variants are sequencing errors or artefacts.

Comparison of four bioinformatic data processing pipelines showed low concordance when identifying discordant variants

To ensure a comprehensive search for real *de novo* mutations, three additional bioinformatic data processing pipelines were implemented on the raw WGS sequencing data (pipelines 2, 3 and 4 as described in Table 3). The discordant variants identified by each pipeline were overlapped for each twin/triplet pair, to identify those most likely to represent real *de novo* mutations. Pipelines 1 and 2 were performed by sequencing provider 1, and pipelines 3 and 4 were performed by sequencing provider 2. Pipelines 1 and 3 each utilised different versions of the Burrows Wheeler Aligner (BWA) [22, 23] and GATK variant calling [21] (BWA-GATK), while pipelines 2 and 4 utilised different version of Isaac [24] alignment and variant calling. Pipelines 1 and 2

(implemented by sequencing provider 1) utilised two-sample VCFs that originated from the aforementioned 850-sample joint-called VCF for the identification of discordant variants, whereas pipelines 3 and 4 (implemented by sequencing provider 2) used two-sample VCFs that were not joint-called. Pipelines 3 and 4 applied a PASS filter prior to discordant variant identification, while the PASS filter was applied to pipelines 1 and 2 following discordant variant identification.

The number of discordant variants identified between ALS affected patients and their unaffected co-twin/triplet varied substantially between the four datasets (using different alignment and variant calling tools) as shown in Table 4.

Table 4. Summary of the discordant variants identified by each bioinformatic processing pipeline for each twin/triplet pair.

		Pipeline 1	**Pipeline 2**	**Pipeline 3**	**Pipeline 4**
Female *SOD1* triplet set	**Pairing A**	12,240 (9,054 SNPs & 3,186 indels)	33,430 (2,506 SNPs & 30,924 indels)	1,947 (1,493 SNPs & 454 indels)	635 (68 SNPs & 567 indels)
	Pairing B	14,097 (9,929 SNPs & 4,168 indels)	15,577 (1,452 SNPs & 14,125 indels)	2,010 (1,534 SNPs & 476 indels)	1,088 (106 SNPs &. 982 indels)
Male *C9orf72* twin set		55,132 (18,759 SNPs & 36,373 indels)	157,012 (3,284 SNPs & 153,729 indels)	6,358 (2,604 SNPs & 3,754 indels)	7,441 (201 SNPs & 7,240 indels)
Female SALS twin set		18,599 (14,160 SNPs & 4,439 indels)	37,226 (2,804 SNPs & 34,422 indels)	1,976 (1,496 SNPs & 480 indels)	1,833 (74 SNPs & 1,759 indels)
Male SALS twin set		30,994 (21,926 SNPs & 9,068 indels)	22,755 (3,411 SNPs & 19,344 indels)	2,646 (1,925 SNPs & 721 indels)	2,480 (126 SNPs & 2,354 indels)

Many more discordant variants were identified from data processed by sequencing provider 1 (pipelines 1 and 2) compared to sequencing provider 2 (pipelines 3 and 4).

Inconsistent distribution of discordant variants between SNP and indel variant types

Over 90% of the discordant variants identified by the Isaac based pipelines (pipelines 2 and 4) were insertion/deletion (indel) variants. In contrast, indels accounted for 20-30% of discordant variants in the BWA-GATK based pipelines (pipelines 1 and 3) in all twin/triplet pairs other than the *C9orf72* twins. The *C9orf72* twins underwent Illumina nano-prep WGS, while all other samples underwent PCR-free WGS. For this twin pair, 50-60% of discordant variants identified from BWA-GATK processed datasets were indels.

The majority of discordant variants are found in intergenic regions

The majority (~52-69%) of discordant variants identified using all four processed WGS datasets were intergenic. Intronic variants were the next most abundant (~15-40%), followed by ncRNA (~6-12%) variants. For all four processing pipelines, less than 1% of discordant variants were exonic. Supplementary Figure 1 provides a breakdown of the distribution of discordant variants between the different genomic functional regions.

Limited overlap of discordant variant datasets

Limited overlap was observed between the discordant variant datasets identified using the four bioinformatics processing pipelines. The discordant

303 variants identified using pipeline 1 were unique, in that they showed no
304 overlap with the discordant variants identified by any other pipeline (i.e.
305 overlap was observed between pipelines 2, 3 and 4, but not with 1). The
306 shared discordant variants that were identified by pipelines 2, 3 and 4
307 represented 0.03-3.2% of the total discordant variants from each pipeline. The
308 Isaac pipelines (pipelines 2 and 4) provided the most shared discordant
309 variants. Fig. 2 provides a visual summary of the overlap of the discordant
310 variant datasets identified by the four pipelines for each twin/triplet pair. For all
311 pairs other than the *C9orf72* twins, 58.3% of discordant variants shared by
312 pipelines 2,3 and 4 were SNP variants, and 41.7% were indel variants. For
313 the *C9orf72* twins, 9.9% were SNPs and 90.1% were indels. Comparison of
314 the two Isaac pipelines (pipelines 2 and 4) showed that 94.9% of shared
315 discordant variants were indels.
316
317 **Putative discordant variants lie in genomic regions that were sequenced**
318 **with low confidence**
319 The sequencing platform used here is known to provide low confidence
320 sequence for some genomic regions including those of low complexity. Zook
321 et al. [25] determined the regions of the genome that are reliably “callable”.
322 Here, all putative discordant variants that were identified by pipelines 2, 3 and
323 4 were shown to fall outside the reliably “callable” genome, as did all
324 discordant variants identified by multiple bioinformatics processing pipelines.
325
326 **Results summary**

In summary, no discordant variants were identified by all four processing pipelines. Pipelines 2 and 4, both of which employed Isaac alignment and variant calling tools, showed the most overlapping discordant variant datasets. Notably, pipeline 1, which utilised a newer version of GATK, showed no concordance with other processing pipelines. Of the shared discordant variants from the three other pipelines, all fell within genomic regions that are known to provide low confidence WGS data. For the two *SOD1* triplet pairings, putative discordant variants shared by pipelines 2, 3 and 4 (pairing A, n=4 and pair B, n=12) showed no overlap. Therefore, no informative discordant *de novo* mutations were identified between ALS discordant twins/triplets.

Discussion

No *de novo* mutations were identified in WGS data that might explain the disease discordance in four sets of MZ twins/triplets. This result is consistent with the extremely low mutation rate of early post-zygotic at just 0.04-0.34x10^{-8} [15]. Therefore, it is expected that only rare cases of disease discordant MZ twins will be explained by *de novo* mutations. Indeed, others have also failed to identify *de novo* mutations that cause SALS discordance between MZ twins [9], suggesting that *de novo* mutations between ALS discordant MZ twins are rare. Similarly, the search for *de novo* mutations in MZ twins discordant for other disorders including Chron's disease [26], Nonsyndromic Cleft Lip and Palate [27], Multiple Sclerosis [28] and Systemic Lupus Erythematosus [29] have also been unsuccessful. The absence of *de*

351 *novo* mutations may be explained by technical downfalls of WGS or
352 alternative mechanisms underlying the disease discordance.
353
354 Other possible mechanisms that may underlie disease discordance in MZ
355 twins include more complex structural variants (SVs) such as copy number
356 variants (CNVs), epigenetic modifications, environmental exposure, or a
357 combination thereof. CNVs are of particular interest for ALS, as the most
358 common known cause of disease, pathogenic expansion of the GGGGCC
359 hexanucleotide repeat unit in *C9orf72* [30, 31], is a type of CNV. Also,
360 intermediate length CNVs in the *ATXN2* gene have been associated with
361 increased disease risk [32], further implicating a role of CNVs in the aetiology
362 of ALS. Other neurodegenerative conditions are also known to be caused by
363 CNVs, including the spinocerebellar ataxias [33-35], Kennedy's disease [36]
364 and Huntington's disease [37].
365
366 Epigenetic modifications are potential contributors to the aetiology of ALS.
367 Differential global DNA methylation levels [38, 39] as well as specific
368 differentially methylated sites [39, 40] having been identified between ALS
369 patients and controls. For example, Meltz Steinberg *et al.* [9] determined that
370 five ALS discordant twin pairs had no genetic discordance but were later
371 found to have significantly different DNA methylation patterns, with
372 accelerated epigenetic aging in affected co-twins [41]. Epigenetic
373 modifications may result from the effects of various environmental exposures
374 [42], including those that have been variably associated with ALS risk [43-46].
375 Epigenetic modifications may represent the intermediary mechanism that links

environmental factors to pathogenic disease mechanisms. Given the shared genetic background, the ALS discordant MZ twins provide an opportunity to decipher the environmental contributions to ALS onset, and their potential link with epigenetic changes.

It is possible that a *de novo* mutation that caused disease discordance evaded detection by the WGS strategy used here. Inadequate coverage of the genome may have seen a region that harbours a *de novo* variant not sequenced, or with an insufficient number of mapped reads. Alternatively, sequencing artefacts introduced during library preparation, sequencing or bioinformatics processing may have masked a true *de novo* mutation. Given that discordant variant identification relied on the comparison of WGS data from two individuals, inadequate coverage or sequencing artefacts in either individual could have prevented identification of a true *de novo* mutation.

In order to remove sequencing artefacts that were introduced by bioinformatics processing, four distinct pipelines were employed that used different alignment and variant calling tools. This found that over 98% of discordant variants were uniquely identified by a single pipeline, suggesting that the overwhelming majority of discordant variants were artefacts introduced by bioinformatics processing. Further, the discordant variants that were identified by multiple pipelines all fell within genomic regions that are notorious for providing sequencing artefacts in WGS data. As such, even if these discordant variants were truly represented in the raw sequencing data,

400 they were likely to be sequencing artefacts introduced by errors that arose
401 during library preparation or sequencing.
402
403 Comparisons between the four bioinformatics pipelines revealed that pipeline
404 1 identified a completely distinct set of discordant variants for each twin pair,
405 in that no putative discordant variants identified by pipeline 1 were identified
406 by any of the other three pipelines. Most interestingly, no discordant variants
407 were shared by pipelines 1 and 3, which both employed BWA and GATK
408 processing tools, albeit different versions. This likely reflects differences in the
409 algorithms used by updated versions of these tools. This highlights the
410 caution required when comparing datasets generated using different
411 alignment and variant calling tools (including updated versions) such that
412 variant identifications between such datasets may not be comparable.
413
414 The comparisons here also highlighted important characteristics of indel
415 variant calls. For the three twin/triplet sets that underwent PCR-free WGS, the
416 discordant variants identified using the BWA-GATK processed datasets
417 (pipelines 1 and 3) consisted of ~20-30% indels, which is comparable to the
418 general abundance of indels across the genome at 21% [47]. However, for the
419 Isaac processed datasets (pipelines 2 and 4), indels accounted for more than
420 90% of discordant variants. As these discordant variants are likely to
421 represent sequencing artefacts, this demonstrates the superior utility of GATK
422 in calling indel variants, as has been reported elsewhere [48, 49]. Indels were
423 also more abundant among the discordant variants identified for the *C9orf72*
424 twin pair, representing ~50-60% of BWA-GATK discordant variants and 99%

425 of Isaac discordant variants. WGS for this twin pair included Nano library
426 preparation, which included a PCR amplification step that may have
427 introduced more indel errors than single nucleotide errors to sequencing
428 templates, as reported by others [50]. This was further supported by the fact
429 that indels accounted for 90.1% of discordant variant identifications shared by
430 pipelines 2, 3 and 4 (n=233) for this twin set. For the other four twin/triplet
431 pairs, indels accounted for just 41.6% (total n=36) of discordant variants
432 shared by these three pipelines.
433
434 All putative discordant variants that were identified by multiple processing
435 pipelines fell within genomic regions that are notoriously difficult to accurately
436 sequence by the WGS strategy used here. The 10% of the genome that
437 harbours highly repetitive sequences, or duplicated genomic elements, has
438 consistently been reported to be the source of abundant false positive variant
439 calls from WGS [25, 51-53]. Therefore, putative discordant variants that were
440 identified in these regions are also likely to represent sequencing artefacts,
441 and were therefore not informative.
442

443 ## Conclusions

444 While no real *de novo* mutations were identified in ALS discordant MZ co-
445 twins/triplets, our analyses highlighted the abundance of sequencing artefacts
446 present in WGS datasets, particularly in difficult to sequence genomic regions,
447 and the substantial differences between alignment and variant calling
448 pipelines. Future analyses will need to increase sequencing coverage and
449 depth, and consider alternative mechanisms of disease onset including

epigenetic modifications and/or environmental exposure. Further, we recommend the use of PCR-free WGS wherever possible, and the application of at least two bioinformatic processing pipelines employing different software tools in order to increase the confidence of all variant identifications.

Materials and Methods

Twins and triplets

Three twin sets and one triplet set (described in Table 1; pedigrees provided in Fig. 1) were ascertained from the Molecular Medicine Laboratory at Concord Hospital and the Macquarie University Neurodegenerative Diseases Biobank. All individuals were recruited under informed written consent as approved by the human research ethics committees of Macquarie University and Sydney South West Area Health Service. All participants were of European descent and the affected co-twin/triplet were clinically diagnosed with ALS based on El Escorial criteria [54]. Genomic DNA was extracted from peripheral blood using standard protocols.

All twins were tested for zygosity using existing SNP genotyping data. SNP genotyping was performed for all four twin sets using either the InfiniumCoreExome-24 v1.0 (*SOD1* triplets and *C9orf72* twins) or v1.1 (both female and male SALS twin pairs) microarray. Raw data was processed using GenomeStudio2011 (Illumina) using standard pipelines. All twins/triplets were also screened for known major ALS genes as described by McCann *et al.* [1].

Generation of whole genome sequence data and raw data processing

DNA samples underwent library preparation using the TruSeq PCR free library preparation kit (Illumina, v2.5), except in the case of the *C9orf72* twin set, for whom the TruSeq DNA Nano kit was used (Illumina). Prepared libraries underwent multiplex 150bp paired-end sequencing on an Illumina HiSeq X Ten instrument (Kinghorn Centre for Clinical Genomics, Sydney, Australia). Four separate bioinformatic processing pipelines were applied to raw WGS data, as detailed in Table 3. Two pipelines utilised each of BWA-GATK and Isaac tools. Pipeline 1 was applied in the initial analysis, while the three other pipelines were applied in the extended analysis.

Discordant variant identification

Discordant variants were defined as genomic sites with a called genotype and a coverage score greater than 30, at which the genotype call differed between the ALS affected co-twin/triplet and their unaffected co-twin/triplet. Custom python scripts were run on a two-sample VCF for each twin/triplet pair, to identify variants that were discordant within a twin/triplet pair. When considering pipelines 1 and 2 (Table 3), the two-sample VCFs originated from a larger, joint-called multi-sample VCF totalling 850 individuals. The two-sample VCFs analysed by pipelines 3 and 4 (Table 3) were not joint-called. Analysis of the triplet set was separated into two triplet pairings. That is, triplet analysis A compared the affected triplet with one unaffected triplet, and triplet analysis B compared the affected triplet to the alternate unaffected triplet.

Comparisons of discordant variants

The BCFTools [55] *isec* command was used to compare VCFs of discordant variants between the four different processing pipelines, triplet pairings A and B, as well as those identified by the original WGS and re-sequencing of the female SALS twin set.

Annotation and distribution of discordant variants

Discordant variant VCFs were annotated for RefSeq genes and genomic functional regions using ANNOVAR [56], and were annotated for variant types using the SNPSift [57] *varType* tool. The distribution of discordant variants between the different genomic functional classes (exonic, intergenic, intronic, ncRNA, splicing, upstream/downstream and UTR), and variant types (SNP and indel), were determined using custom R and bash scripts, respectively.

Variant validation

PCR sequencing

Custom primers were designed for each assessed discordant variant with at least 150bp of flanking sequence. Genomic positions, primer sequences and amplification conditions are available on request. Direct sequencing of amplified fragments was performed using Big-Dye terminator sequencing (v3.1, Applied Biosystems). Sequencing primers were generally the same as amplification primers, however in case of poor sequencing chromatograms, internal sequencing primers were required. In some cases, fragment length analysis was utilised to validate indel variants. This was performed using FAM-labelled forward primers in PCR reactions, and subsequent capillary

electrophoresis of amplified products on an ABI 3730XL sequencer (Macrogen, Korea).

SNP microarray genotyping

SNP genotyping data was generated as described above for zygosity testing. Custom bash and R scripts were used to determine the identity of any WGS discordant variants, from any twin set, that were genotyped by the InfiniumCoreExome-24 v1.0/v1.1 microarrays. The associated genotype data for these SNPs was then extracted and manually analysed in R to determine any discordance within twin/triplet pairs.

Repeat whole genome sequencing

WGS was repeated for the female SALS twin set by Macrogen (Korea). Libraries were prepared using TruSeq PCR free (Illumina) kits and 150bp paired-end sequencing was performed on an Illumina Hiseq Xten instrument. The raw data was processed by Macrogen using Isaac [24] and the corresponding best practices. Discordant variants were identified using the same methods as described above. The BCFtools [55] *isec* command was used to compare the discordant variants identified using the original and re-sequenced WGS data for this twin pair.

Genomic location of discordant variants

To determine whether discordant variants fell within reliably callable regions of the genome, the BCFTools [55] *view* command was used in conjunction with the *regions file* option. Confidently callable regions were defined as those

548 reported by Zook *et al.* [25], the genomic coordinates for which were obtained
549 from [58].
550
551 ## List of Abbreviations
552 ALS: amyotrophic lateral sclerosis; MND: motor neuron disease; FALS:
553 familial amyotrophic lateral sclerosis; SALS: sporadic amyotrophic
554 lateral sclerosis; FTD: frontotemporal dementia; MZ: monozygotic; DZ:
555 dizygotic; NGS: next-generation sequencing; WGS: whole-genome
556 sequencing; SNP: single nucleotide polymorphism; DNA: deoxyribose
557 nucleic acid; PCR: polymerase chain reaction; VCF: variant call file;
558 GATK: genome analysis toolkit; ncRNA: non-coding RNA; UTR:
559 untranslated region; BWA; Burrows Wheeler Alignment; indel:
560 insertion/deletion; SV: structural variation; CNV: copy number variant.
561
562 ## Declarations
563 ### Ethics approval and consent to participate
564 All participants were recruited under informed written consent as approved by
565 the human research ethics committees of Macquarie University and Sydney
566 South West Area Health Service.
567
568 ### Consent for publication
569 Not applicable.
570
571 ### Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the Motor Neurone Disease Research Institute of Australia (grant to KW, PhD top-up scholarship to EM), National Health and Medical Research Council of Australia (grant 1095215 to IB, fellowship 1092023 to KW) and Macquarie University (grant to KW).

The funding bodies did not play a role in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Authors' contributions

EM designed all analyses, performed all bioinformatics analyses of discordant variants, conducted direct sequencing validation experiments, designed all primer sequences, completed all statistical analyses and wrote the manuscript. NT modified and ran bioinformatics scripts used to identify discordant variants and contributed to study design. DB wrote the original bioinformatics scripts used to identify discordant variants. NG performed direct sequencing validation experiments. IB provided intellectual input and supervised the study. KW conceptualised the project, contributed to study design and performed SNP microarray validation. All authors read and approved the final manuscript.

Acknowledgements

We thank Carolyn Cecere, Lorel Adams and Ashley Crook for their assistance in compiling family information, and Elisa Cachia and Sarah Furlong for providing patient materials, clinical and technical assistance.

Figure legends

Fig 1. ALS discordant twin/triplet set pedigrees. Pedigrees for four sets of ALS discordant twins/triplets, with gene mutations indicated. Circles represent females and squares represent males. Filled shapes indicate ALS, open shapes with a dot indicate unaffected mutation carriers and open shapes are unaffected non-carriers. Horizontal lines between twins/triplets indicate monozygosity.

Fig 2. Overlap of discordant variants identified by four different bioinformatics processing pipelines, for each twin/triplet pair. Venn diagrams of the discordant variants identified by the four different bioinformatics processing pipelines described in Table 3. The letters A, B, C and D correspond to pipelines 1, 2, 3 and 4, respectively.

Additional files

Additional file 1. Supplementary information. Figure S1. Distribution of discordant variants between the different genomic functional regions. Stacked bar charts illustrating the distribution of discordant variants between the different genomic functional regions (exonic, intergenic, intronic, non-coding RNA (ncRNA), splicing, upstream/downstream and untranslated region (UTR)). (PDF 3.7MB)

References

1. McCann EP, Williams KL, Fifita JA, Tarr IS, O'Connor J, Rowe DB, Nicholson GA, Blair IP. The genotype-phenotype landscape of familial amyotrophic lateral sclerosis in Australia. Clin Genet. 2017;92(3):259-266.
2. Williams KL, Fifita JA, Vucic S, Durnall JC, Kiernan MC, Blair IP, Nicholson GA. Pathophysiological insights into ALS with C9ORF72 expansions. J of Neurol, Neurosurg & Psychiatry. 2013;84(8):931-5.
3. Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis. Nat Rev Neurol. 2014;10(11):661-70.
4. Tiryaki E, Horak HA. ALS and other motor neuron diseases. Continuum (Minneap Minn). 2014;20(5 Peripheral Nervous System Disorders):1185-207.
5. Ringholz GM, Appel SH, Bradshaw M, Cooke NA, Mosnik DM, Schulz PE. Prevalence and patterns of cognitive impairment in sporadic ALS. Neurology. 2005;65(4):586-90.
6. Phukan J, Elamin M, Bede P, Jordan N, Gallagher L, Byrne S, Lynch C, Pender N, Hardiman O. The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. J Neurol Neurosurg Psychiatry. 2012;83(1):102-8.
7. Montuschi A, Iazzolino B, Calvo A, Moglia C, Lopiano L, Restagno G, Brunetti M, Ossola I, Lo Presti A, Cammarosano S, et al. Cognitive correlates in amyotrophic lateral sclerosis: a population-based study in Italy. J Neurol Neurosurg Psychiatry. 2015;86(2):168-73.

8. Graham AJ, Macdonald AM, Hawkes CH. British motor neuron disease twin study. J Neurol Neurosurg Psychiatry. 1997;62(6):562-9.
9. Meltz Steinberg K, Nicholas TJ, Koboldt DC, Yu B, Mardis E, Pamphlett R. Whole genome analyses reveal no pathogenetic single nucleotide or structural differences between monozygotic twins discordant for amyotrophic lateral sclerosis. Amyotroph Lateral Scler Frontotemporal Degener. 2015;16(5-6):385-92.
10. Al-Chalabi A, Fang F, Hanby MF, Leigh PN, Shaw CE, Ye W, Rijsdijk F. An estimate of amyotrophic lateral sclerosis heritability using twin data. J Neurol Neurosurg Psychiatry. 2010;81(12):1324-6.
11. Xi Z, Yunusova Y, van Blitterswijk M, Dib S, Ghani M, Moreno D, Sato C, Liang Y, Singleton A, Robertson J, et al. Identical twins with the C9orf72 repeat expansion are discordant for ALS. Neurology. 2014;83(16):1476-8.
12. Boomsma DI. Twin, association and current "omics" studies. J Matern Fetal Neonatal Med. 2013;26 Suppl 2:9-12.
13. van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Rev Genet. 2012;13(9):640-53.
14. Zwijnenburg PJ, Meijers-Heijboer H, Boomsma DI. Identical but not the same: the value of discordant monozygotic twins in genetic research. Am J Med Genet B Neuropsychiatr Genet. 2010;153b(6):1134-49.
15. Dal GM, Erguner B, Sagiroglu MS, Yuksel B, Onat OE, Alkan C, Ozcelik T. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. J Med Genet. 2014;51(7):455-9.

16. Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, de Lima RL, Daack-Hirsch S, Sander A, et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat Genet. 2002;32(2):285-9.
17. Castellani CA, Melka MG, Gui JL, Gallo AJ, O'Reilly RL, Singh SM. Post-zygotic genomic changes in glutamate and dopamine pathway genes may explain discordance of monozygotic twins for schizophrenia. Clin Transl Med. 2017;6(1):43.
18. Reble E, Castellani CA, Melka MG, O'Reilly R, Singh SM. VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. Psychiatr Genet. 2017;27(2):62-70.
19. Vogt J, Kohlhase J, Morlot S, Kluwe L, Mautner VF, Cooper DN, Kehrer-Sawatzki H. Monozygotic twins discordant for neurofibromatosis type 1 due to a postzygotic NF1 gene mutation. Hum Mutat. 2011;32(6):E2134-47.
20. Robertson SP, Jenkins ZA, Morgan T, Ades L, Aftimos S, Boute O, Fiskerstrand T, Garcia-Minaur S, Grix A, Green A, et al. Frontometaphyseal dysplasia: mutations in FLNA and phenotypic diversity. Am J Med Genet A. 2006;140(16):1726-36.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
22. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.

695 23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G,
696 Abecasis G, Durbin R. The Sequence Alignment/Map format and
697 SAMtools. Bioinformatics. 2009;25(16):2078-9.
698 24. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies
699 EH, Chuang HY, Kallberg M, Kumar SA, Liao A, et al. Isaac: ultra-fast
700 whole-genome secondary analysis on Illumina sequencing platforms.
701 Bioinformatics. 2013;29(16):2041-3.
702 25. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit
703 M. Integrating human sequence data sets provides a resource of
704 benchmark SNP and indel genotype calls. Nat Biotechnol.
705 2014;32(3):246-51.
706 26. Petersen BS, Spehlmann ME, Raedler A, Stade B, Thomsen I,
707 Rabionet R, Rosenstiel P, Schreiber S, Franke A. Whole genome and
708 exome sequencing of monozygotic twins discordant for Crohn's
709 disease. BMC Genomics. 2014;15:564.
710 27. Mansilla MA, Kimani J, Mitchell LE, Christensen K, Boomsma DI,
711 Daack-Hirsch S, Nepomucena B, Wyszynski DF, Felix TM, Martin NG,
712 Murray JC. Discordant MZ twins with cleft lip and palate: a model for
713 identifying genes in complex traits. Twin Res Hum Genet.
714 2005;8(1):39-46.
715 28. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P,
716 Khrebtukova I, Miller NA, Zhang L, Farmer AD, Bell CJ, Kim RW, et al.
717 Genome, epigenome and RNA sequences of monozygotic twins
718 discordant for multiple sclerosis. Nature. 2010;464(7293):1351-6.

29. Furukawa H, Oka S, Matsui T, Hashimoto A, Arinuma Y, Komiya A, Fukui N, Tsuchiya N, Tohma S. Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. Hum Immunol. 2013;74(2):170-5.
30. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011;72(2):245-56.
31. Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011;72(2):257-68.
32. Elden AC, Kim HJ, Hart MP, Chen-Plotkin AS, Johnson BS, Fang X, Armakola M, Geser F, Greene R, Lu MM, et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature. 2010;466(7310):1069-75.
33. Banfi S, Servadio A, Chung MY, Kwiatkowski TJ, Jr., McCall AE, Duvick LA, Shen Y, Roth EJ, Orr HT, Zoghbi HY. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. Nat Genet. 1994;7(4):513-20.
34. Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LP, Zoghbi HY. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993;4(3):221-6.

744 35. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-
745 Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A, et al.
746 Moderate expansion of a normally biallelic trinucleotide repeat in
747 spinocerebellar ataxia type 2. Nat Genet. 1996;14(3):269-76.
748 36. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH.
749 Androgen receptor gene mutations in X-linked spinal and bulbar
750 muscular atrophy. Nature. 1991;352(6330):77-9.
751 37. A novel gene containing a trinucleotide repeat that is expanded and
752 unstable on Huntington's disease chromosomes. The Huntington's
753 Disease Collaborative Research Group. Cell. 1993;72(6):971-83.
754 38. Chestnut BA, Chang Q, Price A, Lesuisse C, Wong M, Martin LJ.
755 Epigenetic regulation of motor neuron cell death through DNA
756 methylation. J Neurosci. 2011;31(46):16619-36.
757 39. Figueroa-Romero C, Hur J, Bender DE, Delaney CE, Cataldo MD,
758 Smith AL, Yung R, Ruden DM, Callaghan BC, Feldman EL.
759 Identification of epigenetically altered genes in sporadic amyotrophic
760 lateral sclerosis. PLoS One. 2012;7(12):e52672.
761 40. Morahan JM, Yu B, Trent RJ, Pamphlett R. A genome-wide analysis of
762 brain DNA methylation identifies new candidate genes for sporadic
763 amyotrophic lateral sclerosis. Amyotroph Lateral Scler. 2009;10(5-
764 6):418-29.
765 41. Young PE, Kum Jew S, Buckland ME, Pamphlett R, Suter CM.
766 Epigenetic differences between monozygotic twins discordant for
767 amyotrophic lateral sclerosis (ALS) provide clues to disease
768 pathogenesis. PLoS One. 2017;12(8):e0182638.

42. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007;8(4):253-62.
43. Bozzoni V, Pansarasa O, Diamanti L, Nosari G, Cereda C, Ceroni M. Amyotrophic lateral sclerosis and environmental factors. Funct Neurol. 2016;31(1):7-19.
44. Ingre C, Roos PM, Piehl F, Kamel F, Fang F. Risk factors for amyotrophic lateral sclerosis. Clin Epidemiol. 2015;7:181-93.
45. Oskarsson B, Horton DK, Mitsumoto H. Potential Environmental Factors in Amyotrophic Lateral Sclerosis. Neurol Clin. 2015;33(4):877-88.
46. Trojsi F, Monsurro MR, Tedeschi G. Exposure to environmental toxicants and pathogenesis of amyotrophic lateral sclerosis: state of the art and research perspectives. Int J Mol Sci. 2013;14(8):15286-311.
47. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010;19(R2):R131-6.
48. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. PLoS One. 2015;10(11):e0143199.
49. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.
50. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843-51.

51. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al. Comprehensive variation discovery in single human genomes. Nat Genet. 2014;46(12):1350-5.
52. Laurie S, Fernandez-Callejo M, Marco-Sola S, Trotta J- R, Camps J, Chacón A, Espinosa A, Gut M, Gut I, Heath S, Beltran S. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. Hum Mutat. 2016;37(12):1263-71.
53. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5(3):28.
54. Brooks BR, Miller RG, Swash M, Munsat TL. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord. 2000;1(5):293-9.
55. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.
56. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
57. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3:35.

818 58. Genome in a Bottle Consortium website. ftp://ftp-
819 trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/.
820 Accessed 9 November 2018.
821

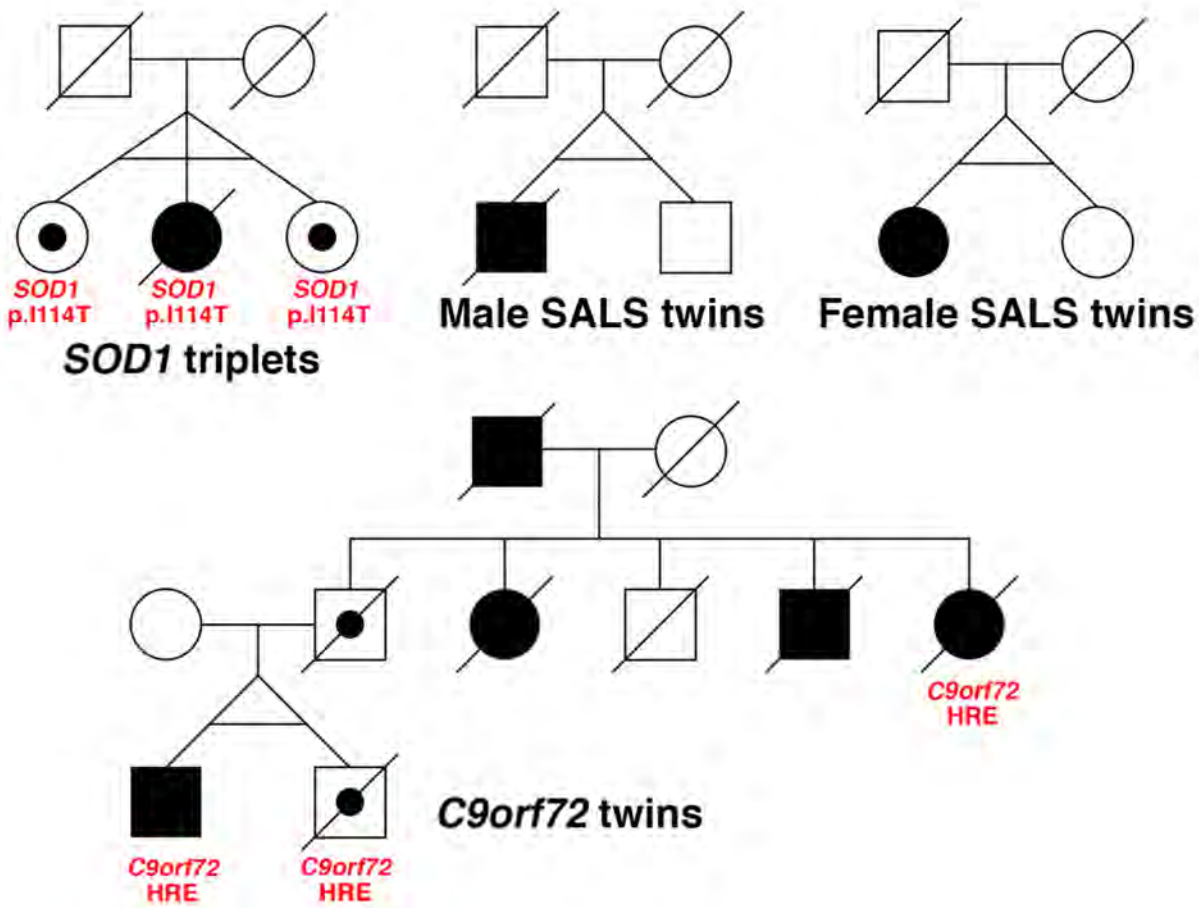


Fig 1. ALS discordant twin/triplet set pedigrees. Pedigrees for four sets of ALS discordant twins/triplets, with gene mutations indicated. Circles represent females and squares represent males. Filled shapes indicate ALS, open shapes with a dot indicate unaffected mutation carriers and open shapes are unaffected non-carriers. Horizontal lines between twins/triplets indicate monozygosity.

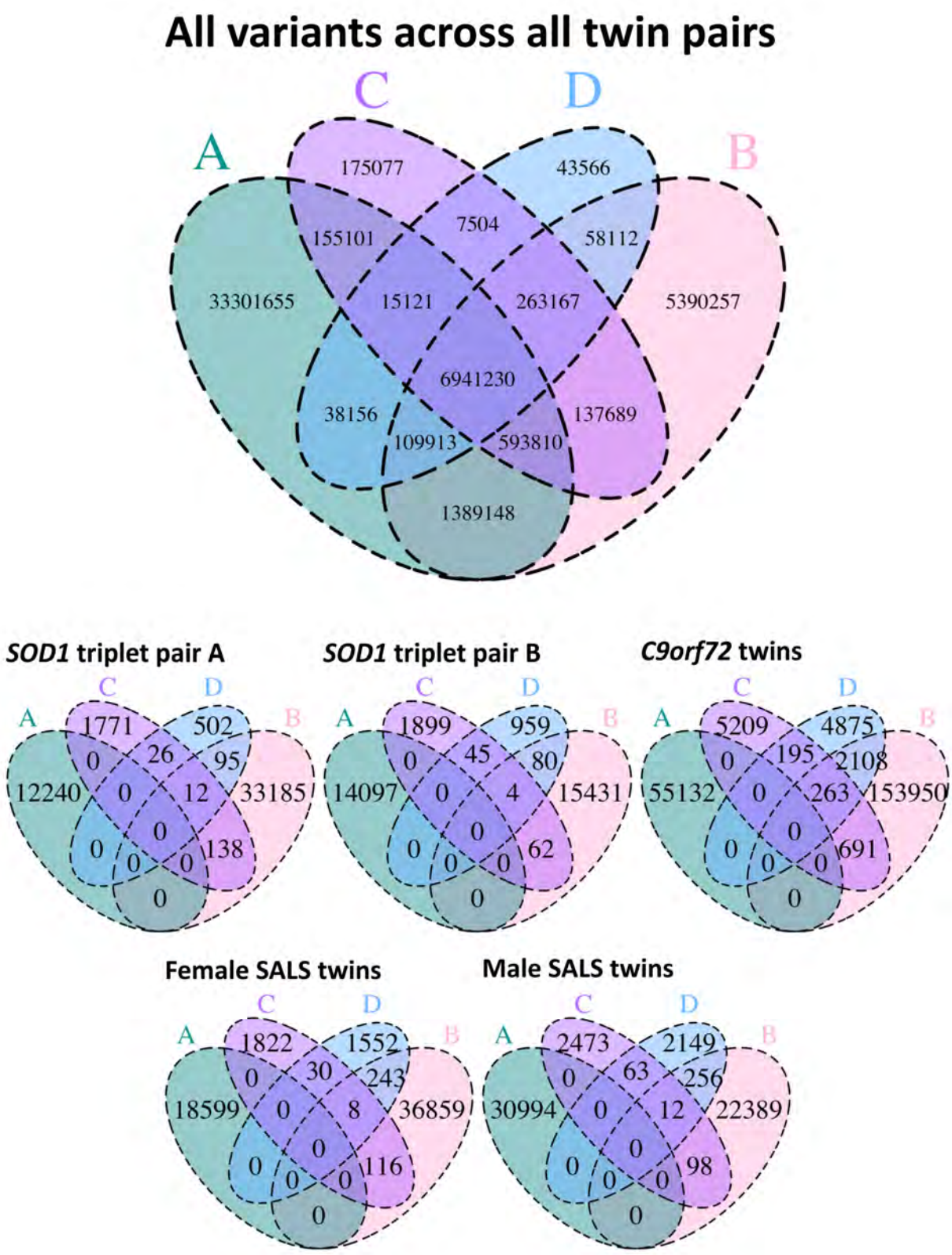


Fig 2. Venn diagrams of the overlap of variants identified by the four different bioinformatics processing pipelines. The top panel shows the overlap of all variants identified across all 11 twins/triplets. The lower panels show the overlap of discordant variants identified for each twin/triplet pair, as described in Table 3. The letters A, B, C and D correspond to pipelines 1, 2, 3 and 4, respectively.

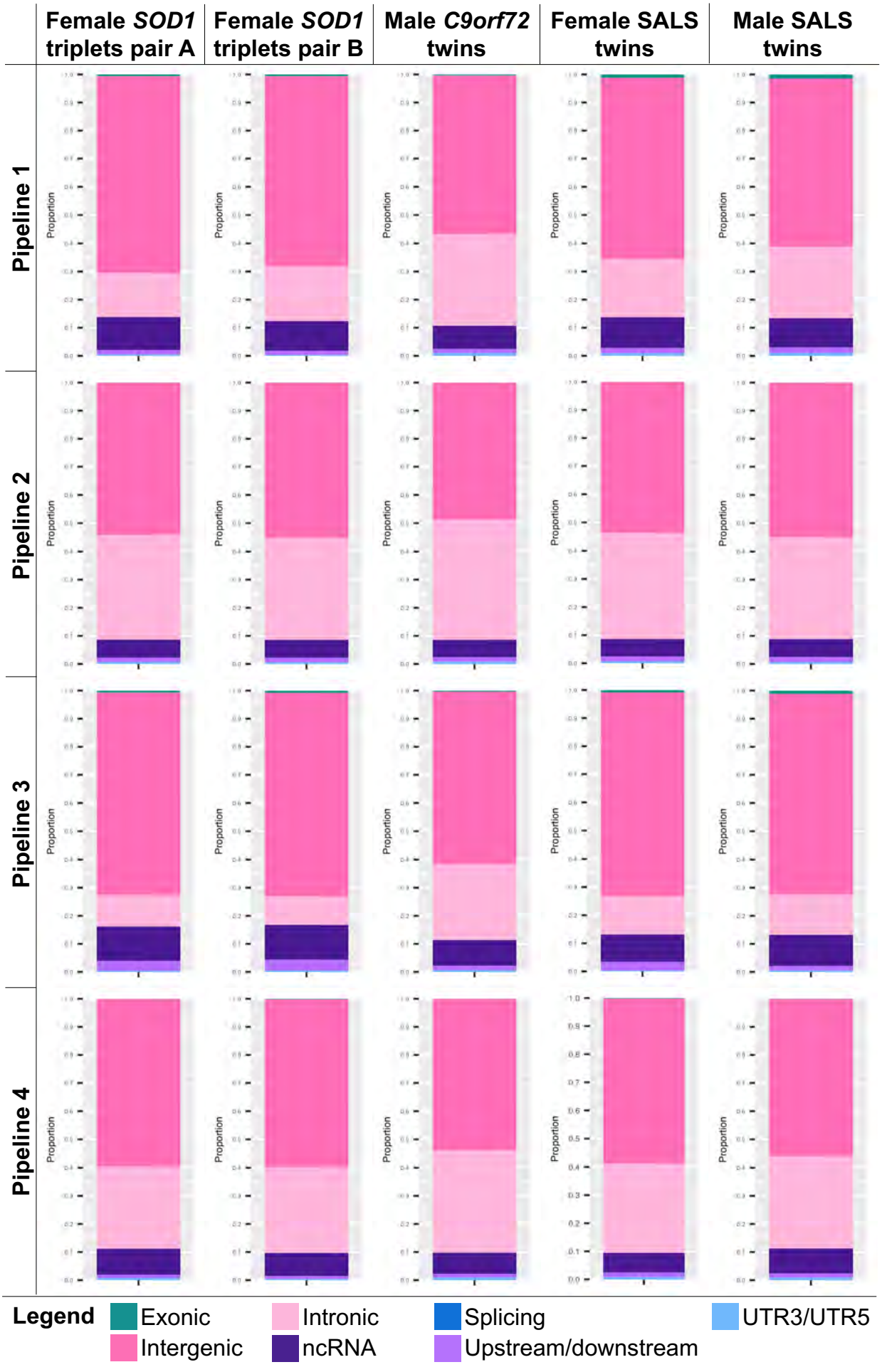


Figure S1. Distribution of discordant variants between the different genomic functional regions. Stacked bar charts illustrating the distribution of discordant variants between the different genomic functional regions.of exonic, intergenic, intronic, non-coding RNA (ncRNA), splicing, upstream/downstream and untranslated region (UTR).

7.2.2 Co-authored Manuscript A4 – Epigenetic and transcriptomic analysis of ALS-discordant monozygotic twin/triplet pairs

In addition to genomic variants, a screen was performed for epigenetic modifications that may contribute to disease discordance between ALS-discordant MZ twin/triplet pairs. The study described in Manuscript A4 (Appendix A.5.4) aimed to characterise DNA methylation and transcriptomic profile differences between ALS patients and their unaffected co-twins/triplets. The hypothesis was that epigenetic modifications and/or differential gene expression may act to modify the phenotypic presentation of ALS between disease-discordant co-twins/triplets. Epigenetic modifications are dynamic and include DNA methylation, histone acetylation and chromatin modelling, among others. Epigenetic modification of gene expression has been proposed as the mechanism by which the exposure to environmental risk factors can influence the molecular mechanism of disease. In part, the rationale behind an epigenetic contribution to ALS stems from the accumulation of these marks over time, which fits well with the late onset of disease and phenotypic variability between patients.

Here, ALS affected co-twins were shown to have significantly increased epigenetic ages compared to their unaffected co-twin, which supports the findings of a similar study perfromed by Young et al. (2017). A total of 59 differentially methylated sites were identified across all four twin/triplet sets. Two of these sites, in the genes *C8orf46* and *RAD9B*, also showed significant differential methylation in a case-control cohort of 650 SALS patients and 539 control individuals. Unfortunately, the set of 59 differentially methylated probes did not have sufficient power to discriminate between ALS affected and unaffected co-twins/triplets, or SALS patients and controls. Analysis of within-twin-set differential methylation identified two probes and 13 genes that showed differential methylation in more than one disease-discordant twin set, that may represent potential risk or modifier markers of ALS. Interestingly, transcriptomic analysis showed that two genes previously implicated in ALS, *CCNF* and *CCS*, were downregulated in the ALS affected twin/triplet compared with their unaffected co-twin/triplet. The integration of methylomic and transcriptomic profiling data highlighted twelve genes that were both differentially methylated and expressed. This study implicated an epigenetic contribution to disease in the ALS-discordant twins/triplet sets. It is likely that in the broader ALS patient population, and possibly within these twin sets, that these epigenetic modifiers act together with environmental risk factors and genetic variation to influence the onset and/or progression of disease.

Author contributions
IT and KW conceived and designed the study with input from SC, NW and IB. IT, EM (the candidate), BB, TP, NT, KZ, QZ, Z-HZ, DB and KW performed the experiments/data analysis and interpretation. DR and GN collected clinical information and samples. IT and KW wrote the manuscript with input from IB. All authors read and approved the final manuscript.

7.3 Discussion

In this Chapter, the utility of ALS-discordant monozygotic twins has been explored for the identification of nucleotide level variation or epigenetic and transcriptomic alterations that contribute to the aetiology of ALS. While the genomic investigation was unsuccessful in identifying novel genetic causes or modifiers of ALS, this was not altogether surprising. The post-zygotic *de novo* mutation rate is low at just 0.04-0.34$\times10^{-8}$ (Dal et al., 2014), and many others have also failed to identify nucleotide level variation between disease-discordant MZ twins. This includes investigations of diseases such as Crohn's disease (Petersen et al., 2014), non-syndromic cleft lip and palate (Mansilla et al., 2005), multiple sclerosis (Baranzini et al., 2010) and systemic lupus erythematosus (Furukawa et al., 2013). Most notably, Meltz Steinberg et al. (2015), conducted a similar study to that reported in Manuscript III, in which five SALS twin pairs underwent WGS for discordant variant identification. They too were unable to identify any nucleotide level variation explaining disease discordance.

Though we were unable to identify any *de novo* nucleotide mutations contributing to the aetiology of ALS in these four discordant twin/triplet sets, it is possible that our analysis was underpowered to detect such mutations. This could be the result of tissue specificity or inadequate WGS coverage. As sequencing was performed using DNA derived from peripheral blood, and ALS affects neuronal tissue, a somatic mutation that occurred further down the cell lineage may have been missed. Indeed, it has been reported that substantial genetic variation exists between different tissue types (O'Huallachain et al., 2012), and as such, it may be that a disease causing or modifying mutation may be confined to the disease-affected tissue. Even if neuronal tissue was available for sequencing, it is likely that such a mutation would still be missed, as many somatic mutations are only present in a very small number of cells, causing these to go undetected using standard coverage WGS (Ye et al., 2013). Additionally,

it has been suggested that 100X sequencing is necessary to achieve 100% coverage of the human genome, so that at least 20 reads are mapped to each nucleotide position (Meienberg et al., 2016). Here we have used 30X sequencing, which while commonly employed and accepted to strike an acceptable balance between cost and sensitivity (Lohmann and Klein, 2014), may be inadequate to detect rare variants or those falling in difficult to sequence genomic regions. Further, variants in highly polymorphic and low complexity genomic regions are not well represented by NGS read data (Li, 2014; Tian et al., 2016; Weisenfeld et al., 2014), therefore *de novo* mutations in such regions may not have been detected here.

The large number of apparently discordant variants initially identified between ALS-discordant co-twins/triplets using WGS is in fact consistent with other twin studies. For example, Illumina WGS of two MZ twin pairs by Ye et al. (2013) identified approximately 30,000 discordant variants in each pair, while Complete Genomics sequencing identified approximately 14,000. They found that most of these discordant variants were sequencing platform artefacts, as the intersection of discordant variants identified by the two platforms was just 13 and 17, of which just eight were validated by Sanger sequencing. In another study, WES of nine monozygotic twin pairs identified a total of almost 7,000 discordant variants, which were all later found to be artefacts (Zhang et al., 2016). Given the scope of WES to WGS, these discordant variant numbers are comparable to those identified in Manuscript III and by Ye et al. (2013) using WGS.

Validation efforts suggested that the discordant variants identified here were also largely artefacts. After using SNP microarrays and gold standard Sanger sequencing to directly validate individual genotypes, and failing to validate a single variant as truly discordant, it was concluded that the majority of the putative discordant variants were artefacts of the WGS pipeline. Given the impracticality of directly sequencing tens of thousands of variants, WGS was repeated for one twin pair, however cost constraints prevented this validation approach in all twin/triplet sets. Failure to replicate a single putative discordant variant in the new WGS dataset affirmed the likelihood that all discordant variants between MZ twins/triplet pairs represented artefacts of the WGS pipeline.

In an effort to use various bioinformatics processing pipelines to identify true *de novo* mutations, it was shown that over 98% of all putative discordant variants were identified by just one of the four processing pipelines, strongly implicating

these variants as sequencing artefacts introduced by bioinformatics processing. Unfortunately, due to time constraints of this candidature, Sanger sequencing of the overlapping variants was not possible, however this validation will be completed prior to submission of Manuscript III for publication. As an alternate approach to assess the likelihood that the overlapping variants were truly discordant between ALS affected and unaffected co-twins/triplets, the genomic context of each was determined. This showed that all overlapping discordant variants fell outside of the confidently "callable" genome (as defined by Zook et al. (2014)), and rather fell within the 10% of the genome which is notoriously difficult to accurately genotype and enriched for sequencing artefacts (discussed further in Chapter 8, Section 8.3.3 Laurie et al., 2016; Telenti et al., 2016; Weisenfeld et al., 2014; Zook et al., 2014). This suggested that these overlapping putative discordant variants were also most likely sequencing artefacts rather than true *de novo* mutations.

It was expected that far better concordance would be apparent between the discordant variants identified by pipelines 1 and 3, and pipelines 2 and 4, as the same basic alignment and variant calling tools were shared by these pipelines, being BWA-GATK and Isaac, respectively. However, the most notable difference between the number of discordant variants identified between the four different processing pipelines was that datasets processed by service provider 1 reported between six- and 52-fold more discordant variants than those processed by service provider 2. This increase is likely attributable to two major factors. Firstly, the inconsistencies in applying the PASS filter to WGS data. Service provider 1 did not apply the PASS filter in their processing pipelines (1 and 2), while service provider 2 (pipelines 3 and 4) did. This therefore increased the number of low quality variants (and therefore sequencing artefacts) in pipelines 1 and 2, and subsequently the number of putative discordant variants identified by analysis of these datasets. In an effort to rectify this inconsistency, the PASS filter was applied to the discordant variants identified by pipelines 1 and 2, however this failed to substantially reduce discordant variant identifications. This is attributable to the fact that when retrospectively applying the PASS filter, only one co-twin/triplet was required to possess a PASS annotation, while when applying the PASS filter prior to discordant variant identification, both co-twin/triplets were required to possess a PASS annotation, which substantially increased filtering stringency. Secondly, the differences between the joint-calling methodology applied by the two service providers is likely to have contributed to this discrepancy, as multiple-sample calling is well established to increase specificity and the number of variant calls, but incidentally may also increase the number of

false positive variant calls (Liu et al., 2013). Given that the dataset processed by service provider 1 was joint-called using a cohort totalling 850 samples, and that by service provider two was not joint-called, an inflation of false positive variants in data processed by service provider 1 is to be expected.

Indels are notoriously difficult to call (Pabinger et al., 2014), as evidenced by very low concordance between indel calls between various variant calling tools (O'Rawe et al., 2013). Therefore, we would expect that there would be an over-representation of indels among the discordant variant calls. It has been reported that indels account for approximately 21% of nucleotide level variants, while SNPs make up the balance (Mullaney et al., 2010). Surprisingly, this ratio is reflected by the discordant variants reported by the BWA-GATK processed datasets. However, the Isaac processed datasets report up to >90% of discordant variants as indels. These results likely reflect the difficulties encountered by Isaac in calling indel variants, and the superior ability of GATK to call these variant types, which has been reported previously (Field et al., 2015; Hwang et al., 2015).

While we have established that the vast majority of discordant variants between ALS-discordant co-twins/triplets were attributable to errors introduced by bioinformatics processing, it is important to note that as a multi-component process, WGS artefacts may have been introduced at any point of the WGS pipeline. In addition to bioinformatic errors, WGS artefacts may also result from PCR amplification or sequencing errors. A detailed discussion of sequencing artefacts will be provided in Chapter 8, Section 8.3.3. Therefore, the discordant variants identified by multiple processing pipelines may actually be artefacts introduced by alternate sources of error in the WGS pipeline. It is also possible that sequencing artefacts may have masked the identification of true *de novo* mutations between co-twins/triplets, by introducing concordant genotypes. However, the likelihood of such an event is exceptionally low.

Across all four processing pipelines, the *C9orf72* twin pair had between two and 10-fold more discordant variants than any other twin/triplet pair. This is to be expected, as this pair underwent PCR amplification prior to WGS, while all other samples did not. PCR amplification is notorious for introducing sequencing artefacts (Aird et al., 2011; Brockman et al., 2008; Li, 2014; Meienberg et al., 2016). Genomic regions of high GC content, and those with poly(A) stretches or AT dinucleotide repeats are frequently incorrectly amplified due to polymerase errors (Aird et al., 2011; Brockman et al., 2008; Meienberg et al., 2016). Additionally, PCR amplification

can introduce allele bias, as multiple reads originating from a single template fragment may be produced by sequencing, which can lead to errors in variant calling statistics (Pabinger et al., 2014). As such, it may be that a substantial proportion of the 233 discordant variants in this twin pair, consistently identified by three of the four variant callers, were truly present in the template sequence, but were indeed artefacts introduced by PCR amplification. Gold standard Sanger sequencing of the original co-twin DNA samples and WGS template library would be required to determine whether this was indeed the case.

As no nucleotide level variants were identified as contributors to the cause or differential presentation of ALS in these MZ co-twins/triplets in Manuscript III, other molecular perturbations may be responsible for disease discordance. These include, structural variants (SVs) such as copy number variants (CNVs) and repeat expansions, epigenetic modifications such as DNA methylation or histone modifications, or environmental exposures. It is possible that a combination and/or accumulation of these factors is required to trigger disease onset, as is suggested by the multi-step hypothesis (Al-Chalabi et al., 2014; Chio et al., 2018) discussed in Chapter 8, Section 8.2.1.6.

Structural variants, particularly repeat expansions and CNVs, represent a likely alternative genetic alteration to underlie disease discordance between ALS-discordant co-twins/triplets. The disease-discordant MZ twin model has previously had success in identifying pathogenic SVs (Dahoun et al., 2008; Ramsey et al., 2012; Razzaghian et al., 2010), including CNVs and repeat expansions (Breckpot et al., 2012; Bruder et al., 2008). Repeat expansions, which are in essence a type of CNV, have been implicated as contributing to the aetiology of ALS, with pathogenic expansion of a hexanucleotide repeat unit in the *C9orf72* gene being the most common known cause of disease (DeJesus-Hernandez et al., 2011; Renton et al., 2011), and intermediate length ATXN2 expansions having been shown to increase ALS-risk (Elden et al., 2010). Chapter 8, Section 8.2.1.4 will provide a discussion of the potential broader role of repeat expansions and CNVs in ALS pathogenesis. The planned investigation of CNVs in ALS, including within these twin/triplet sets, is discussed in Chapter 8, Section 8.5.4.

The epigenetic differences identified between co-twins in Manuscript A4 highlighted the involvement of epigenetic and gene expression mechanisms in ALS. However it remains to be seen whether these changes are a cause or consequence of disease. The most robust finding of this manuscript was that of accelerated epigenetic aging of the

ALS affected co-twin/triplet, which has also been demonstrated by previous studies (Young et al., 2017; Zhang et al., 2016). As a degenerative disease, age is a major ALS risk and prognostic factor (Chio et al., 2009a) and increased DNA methylation age has been associated with increased mortality (Christiansen et al., 2016). The analyses presented in Manuscript A4 did not find any changes in global DNA methylation levels, in contrast to other studies (Chestnut et al., 2011; Figueroa-Romero et al., 2012). The identification of decreased expression for the ALS gene *CCNF* in a SALS patient compared to his unaffected co-twin is also very interesting. It has previously been shown that mutations in *CCNF* cause ALS (Williams et al., 2016b), and that there is a significant burden of protein-altering variants in this gene among SALS cases (Williams et al., 2016b). It is yet to be determined whether *CCNF* mutations lead to a loss- or gain-of-function, but the findings presented in Manuscript A4 suggest that reduced *CCNF* expression may play a role in the presentation of ALS.

This epigenetics data suggests that gene expression is an important factor in disease discordance between these four ALS-discordant MZ twin/triplet sets. Therefore, *de novo* mutations that lie within regulatory regions and have potential to impact gene expression, may have a functional effect on disease phenotypes. Large SVs are also likely to affect regulatory sequences, impacting gene expression that may contribute to disease discordance between MZ co-twins/triplets. Indeed, SVs have often been reported as having regulatory functions (Weckselblatt and Rudd, 2015). Together, this supports a more thorough investigation of non-coding regulatory regions in the aetiology of ALS. Ideally, this would involve high coverage WGS and in-depth analysis of SVs.

The lack of discordant *de novo* mutations (Manuscript III), together with evidence of epigenetic differences (Manuscript A4) between the four ALS-discordant monozygotic twin/triplet sets suggests that environmental risk factors may have played a role in triggering the onset or progression of ALS in these cases. As described in Chapter 1, Section 1.3.2, a variety of environmental factors have been proposed to increase the risk of developing ALS. However, these associations are often not supported by follow-up studies. As epigenetic modifications are thought to reflect many environmental exposures, the case may be that a complex interplay between environment, epigenetics, genetics and/or other unknown factors may trigger the onset of ALS. The effects of epigenetic modifications instigated by environmental exposure may also be influenced by genotype, and therefore vary between individuals. This could explain why an association between an environmental

factor and disease in a given population is not replicated in other populations with different ancestral backgrounds. The identical genetic background of MZ twins provides an excellent opportunity to investigate environmental factors influencing the onset or progression of ALS, and the epigenetic changes induced by such exposures.

"If I can live through this, I can do anything."

Fallout Boy - Champion

8 Discussion

8.1 Summary of results

This thesis has presented a comprehensive investigation into the genetic basis of Australian ALS using next-generation sequencing (NGS) data. Numerous novel bioinformatics pipelines were established to effectively analyse and utilise the plethora of genetic information produced by whole-exome (WES) and whole-genome (WGS) sequencing. Application of these strategies, together with various traditional genetic analysis techniques, has furthered our understanding of the genetic spectrum of Australian ALS. Analysis of Australian ALS families showed that 60.8% carry mutations in known ALS genes (Figure 8.1). The recently reported ALS genes, *CHCHD10*, *C21orf2*, *GPX3-TNIP1* and the *hnRNP* genes were also assessed, however none were shown to contribute to the cause of ALS amongst Australian patients. A screen for 54 candidate ALS genes in the Australian patient population identified eight candidate ALS mutations, and 17 candidate variants potentially associated with ALS as disease-risk or protective alleles.

This project also sought to identify novel ALS causal gene mutations in familial ALS. Using WES and WGS data, together with custom bioinformatics strategies and a custom *in silico* variant prioritisation approach, substantial progress was made

towards identifying the causal gene mutation in five small ALS families. Four of these families now have short lists of candidate ALS mutations, with strong evidence supporting a causal role for just five, six, one and 11 of these, from an initial pool of more than 90,000 genetic variants in each family. While no nucleotide level mutations were identified in the fifth family, the combination of linkage analysis with next-generation sequencing facilitated the exclusion of ~86.64% of the genome from harbouring the unidentified causal mutation. Analysis of this family successfully excluded numerous potential candidate mutations, thereby providing vital guidance to future studies to investigate alternate types of genetic variation, including structural variants, as a cause of disease in this family. Similarly, the search for novel ALS genes using WGS data from disease-discordant monozygotic (MZ) twins/triplets suggested that, in peripheral blood, no somatic *de novo* nucleotide level variants were responsible for disease in the affected co-twin/triplet. While this is a negative result, it is highly informative in that it removes early post-zygotic nucleotide level variation as a cause of disease onset in these ALS-discordant twin/triplet sets, so that other causes of disease discordance may be explored.

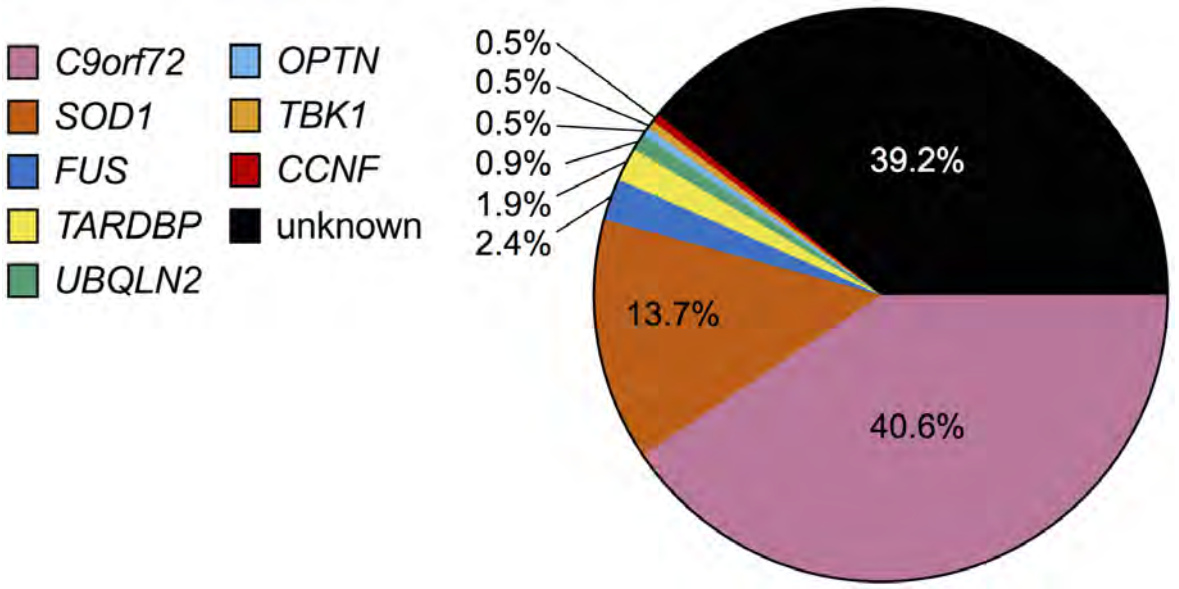


FIGURE 8.1: **The genetic landscape of Australian familial ALS.** Eight genes were found to harbour ALS causal mutations in the Australian FALS cohort of 212 families, accounting for 60.8% of this cohort. Paper I, Section 4.3.1.

8.2 Gene discovery in ALS

ALS is a genetically and phenotypically heterogeneous disease. The data presented in Chapters 4 – 6 demonstrated the immense genetic heterogeneity of ALS, while investigations of the correlations between clinical phenotypes and the known ALS gene mutations in Paper I (Chapter 4, Section 4.3.1) highlighted the phenotypic heterogeneity of disease. Twenty-one distinct mutations, from eight different genes were found to cause disease amongst Australian FALS patients in Paper I (Chapter 4, Section 4.3.1) accounting for 60.8% of the 212 family cohort. Therefore, a large portion of the genetic aetiology of FALS remains to be identified. The genetic landscape of disease is likely to be far greater than the 25 known ALS genes identified to date. This is supported by data presented in this thesis, where the search for ALS causal mutations in five unsolved ALS families (Chapter 6) identified a unique list of candidate mutations for each family. This suggests that the remaining causal familial ALS mutations are likely to be extremely rare variants, perhaps even private to their respective families. Further, candidate gene screening (Chapter 5) identified eight unique candidate mutations in six different genes, each found in a single patient. Combined, these factors act to contribute to the heterogenic nature of ALS.

The reduced penetrance seen in ALS families is a considerable barrier to solving the genetic basis of the remaining 39.2% of Australian ALS families with unknown causal mutations. Inherently, families with a history of ALS and reduced disease penetrance will have less affected family members, and therefore fewer informative DNA samples available for genetic interrogation. This is evidenced by the high proportion of families in the Australian cohort for which DNA samples are only available from the proband patient, or two to three genetically informative family members, including the five families analysed in Chapter 6. In all but one of these five families, traditional genetic linkage analysis was not effective due to the limited availability of DNA samples. Genetic linkage analysis has played an instrumental role in disease gene discovery for over two decades, and in the absence of the genetic mapping information provided by linkage analysis, disease gene discovery is at a considerable disadvantage. Therefore, alternative approaches are necessary to identify ALS causal mutations in these families. While NGS, and particularly WES, has facilitated the discovery of many disease genes, this approach also has limited power when analysing very small families.

Complex statistical tools have recently been developed that seek to apply modified

linkage strategies to NGS datasets. The Pedigree Variant Annotation, Analysis and Search Tool (pVAAST) employs a sequence based model to perform both gene- and variant-based linkage analysis using NGS data from families affected by disease (Hu et al., 2014). The results of linkage analysis are then combined with functional predictions and rare variant case-control association testing, to rank variants based on their likelihood of pathogenicity (Hu et al., 2014). While the authors report that pVAAST is robust to incomplete penetrance and locus heterogeneity (Hu et al., 2014), validation trials of pVAAST performed during this candidature using WES data from Australian families carrying known pathogenic *UBQLN2* and *CCNF* mutations failed to identify these as top-ranked disease genes (data not shown). This suggested that the pVAAST tool was not appropriate for use in small- to medium-sized ALS families with incomplete penetrance, and therefore this avenue of analysis was not further pursued. To successfully and robustly identify disease causal mutations in complex diseases such as ALS, statistical models will need to account for disease prevalence, disease penetrance, natural genic variation and mutation rates (MacArthur et al., 2014). To-date, no such models are available, and substantial methodological development will be required for these to be established (MacArthur et al., 2014).

Increasingly, there is a general acceptance that a genetic underpinning exists for some, if not all, sporadic ALS cases (Al-Chalabi et al., 2017). Indeed, some SALS patients have been found to harbour causal mutations in known ALS genes, including *C9orf72*, *TARDBP*, *FUS*, *SOD1*, *EWSR1*, *CCNF* and *TIA1* (Couthouis et al., 2012; Gu et al., 2018; Sreedharan et al., 2008; van Blitterswijk et al., 2012a; Vance et al., 2009; Williams et al., 2013, 2016b; Yuan et al., 2018; Zhang et al., 2018). In Chapter 4, the prevalence of pathogenic hexanucleotide repeat expansions of *C9orf72* was found to be 2.9% in Australian SALS, while two novel candidate mutations were identified in two separate SALS patients in the *TIA1* gene as part of Chapter 5. Gene mutations causing disease in ALS families with reduced disease penetrance may represent mutations of medium to strong effect, therefore such mutations may also be implicated in the cause of disease in some apparently sporadic patients.

Three different approaches to ALS gene discovery were employed during this project including, candidate gene screening, family-based analysis and analysis of disease-discordant MZ twins/triplets. Each of these approaches have innate advantages and disadvantages. Candidate gene screening is an efficient approach for utilising otherwise uninformative (for gene discovery purposes) data from FALS probands and SALS patients. However, unless replicated in additional cohorts, the pathogenicity

of identified candidate gene mutations or associated variants remains ambiguous. Family-based analysis offers the most robust approach to ALS gene discovery, however the small size of the families used in this project resulted in lists of candidate mutations, and further, the genetic heterogeneity of ALS prevented pooling of family data for disease gene discovery. Lastly, genomic comparisons between ALS-discordant MZ twins is a straight forward and potentially compelling avenue for disease gene discovery. Unfortunately, the disease-discordant monozygotic twin model has a low success rate (Baranzini et al., 2010; Furukawa et al., 2013; Mansilla et al., 2005; Meltz Steinberg et al., 2015; Petersen et al., 2014), given the extremely low frequency of post-zygotic somatic mutations.

8.2.1 Beyond Mendelian disease

While the vast majority of known ALS causal mutations identified to-date are nucleotide level protein-altering variants, the most common cause of disease (hexanucleotide repeat expansion of a GGGGCC repeat in the first intron of *C9orf72*) is a non-coding structural variant. Together with the difficulties encountered by gene discovery efforts in the unsolved ALS families, and the lack of known causes of sporadic disease, this suggests that it is probable that other types of genetic mechanisms may cause ALS. The following sections will provide an overview of such possible genetic mechanisms that may be acting to cause ALS.

8.2.1.1 Oligogenics

Oligogenic inheritance refers to a small number of gene variants acting together to influence a particular phenotype. It represents an intermediary between classical monogenic Mendelian inheritance and complex polygenic inheritance, in which phenotype is determined by one or many gene variants, respectively. The most compelling data supporting an oligogenic basis of ALS was reported by van Blitterswijk et al. (2012a), who identified that five of 97 FALS families carried mutations in two ALS genes (defined as *C9orf72*, *SOD1*, *FUS*, *TARDBP* or *ANG*). Indeed, our own laboratory has also reported two ALS families carrying both a pathogenic expansion of *C9orf72*, and a gene mutation in *ANG* (Williams et al., 2013).

With the widespread adoption of NGS technologies, reports of an oligogenic basis of ALS or ALS/FTD have become more frequent (Keogh et al., 2018; Zou et al., 2017). Apparent oligogenic inheritance in ALS is most frequently reported for patients

carrying a *C9orf72* expansion, the most common known cause of ALS (Nguyen et al., 2018b; Zou et al., 2017). This is interesting given that the *C9orf72* expansion can cause either ALS, FTD or co-morbid disease, thus the influence of additional genetic variation may contribute to the phenotypic manifestation of disease (van Blitterswijk et al., 2012a). Observations by Nguyen et al. (2018b) support this notion, in that patients with a *C9orf72* expansion as well as a mutation in one of the ALS genes *FUS*, *OPTN*, *ANG* or *SOD1*, invariably presented with pure ALS, while those with a *GRN* mutation always presented with pure FTD. Further, many *ANG* mutations have also been reported in an oligogenic state with a more common ALS gene (van Blitterswijk et al., 2012a). This may reflect that *ANG* mutations require oligogenic inheritance with another ALS mutation to cause disease, rather than eliciting a sufficient effect to cause disease in isolation. Further, following the report of *NEK1* as an ALS gene due to an increased burden of loss-of-function variants in this gene among patients (Kenna et al., 2016), Nguyen et al. (2018c) found that over 50% of *NEK1* carriers had an additional mutation in another ALS gene, suggesting a potential oligogenic role for *NEK1* in ALS.

Keogh et al. (2018) reported that 19 of 244 ALS/FTD patients carried more than one variant in an ALS/FTD gene. Eleven of these 19 patients carried a known, or likely pathogenic mutation, and the additional variant in each of these patients was assessed to be benign or likely benign based on the American College of Medical Genetics and Genomics (ACMG; Richards et al., 2015) guidelines. The other eight of these 19 oligogenic cases carried two or more benign/likely benign variants in ALS/FTD genes. Such benign assessments cast doubt over the actual contribution of such variants to the cause of disease. Application of pathogenicity assessment criteria to the variants identified by the aforementioned studies would be worthwhile, as it may weaken the support for an oligogenic basis of ALS. Alternatively, additional ALS gene variants may be acting as phenotypic modifiers of disease rather than causal mutations (Therrien et al., 2016). While there is certainly cause to consider the possibility of an oligogenic basis of ALS, further evidence will be required to determine the extent to which this mechanism contributes to ALS pathogenesis.

The families presented in Chapter 6, as well as the families of the probands in Chapter 5 show reduced disease penetrance that could in part be explained by an oligogenic disease model. Additionally, as each of the four small families from Chapter 6 had lists of candidate mutations, it is possible that more than one of these genetic variants are interacting to cause disease. Given that in three of these four families, multiple candidate mutations were assessed as having a high potential for

pathogenicity using the *in silico* pipeline, this scenario is plausible.

8.2.1.2 Risk genes

As discussed in Chapter 1, Section 1.4.2, the predominance of sporadic disease and high heritability estimates for all forms of ALS suggest that genetic variants of small to moderate effect may contribute to the risk of developing SALS. While genome-wide association studies (GWAS) have previously had limited success in SALS (Renton et al., 2014), more recently, meta-analyses of GWAS data have identified multiple loci as robustly associated with SALS. These include variants in the genes *C21orf2* and *GPX3-TNIP1*, which were genotyped through Australian SALS and control cohorts in Chapter 4 as part of Papers A1 and A2 (Appendices A.5.1 and A.5.2). These loci however did not show association with disease within the Australian population. This is likely attributable to either the smaller sample size of the Australian replication cohort, or the unique genetic landscape of ALS in the Australian patient population (as discussed in Chapters 4 and 5).

The work by van Rheenen et al. (2016) (Paper A1; Appendix A.5.1) also identified *MOBP* and SCFD1 as ALS risk genes, and showed that SALS has a complex genetic architecture with a SNP-based heritability of 8.5%. Rare variants were shown to make a substantial contribution to this heritability, which goes some way to explaining why GWAS (which is based on the principle of "common variant – common disease") has had limited success in identifying risk alleles in SALS cohorts. Further, association analysis was instrumental in the discovery of the pathogenic expansion of *C9orf72* as the most common known cause of ALS (Laaksovirta et al., 2010; Shatunov et al., 2010; van Es et al., 2009). As such, it may be that association-based analyses will also be required to identify other novel ALS causal mutations.

Another possibility is that polygenic inheritance is at play in SALS. Polygenic inheritance refers to instances where a phenotype is mediated by the cumulative effect of numerous genetic variants (Torkamani et al., 2018). Polygenic risk scores have recently emerged as a way of assessing the level of susceptibility an individual has to developing disease, given their genomic profile (Torkamani et al., 2018). Polygenic risk scores for most diseases, including ALS and other neurodegenerative diseases, are likely to be influenced by both rare and common genetic variants (Ibanez et al., 2019; Torkamani et al., 2018). Indeed, polygenic risk scores have previously been described for ALS, as well as FTD and Alzhiemers disease (Hagenaars et al., 2018). With

the increasing availability of NGS data from ALS patients, and the accompanying expansion of patient cohorts with WGS data, polygenic risk scores will be refined and consolidated.

8.2.1.3 Genetic burden

Genetic burden analysis compares the cumulative frequency of qualifying variants (which may be defined based on any criteria, such as non-synonymous variants) in a gene between cases and controls, to determine whether that gene carries a significant burden of genetic variation. The growing availability of NGS data from large cohorts is now making genetic burden analysis possible in ALS, as described in Chapter 1, Section 1.4.2.2. Numerous ALS genes have been implicated through genetic burden testing including *TUBA4A* (Smith et al., 2014), *TBK1* (Cirulli et al., 2015; Freischmidt et al., 2015) and *NEK1* (Cirulli et al., 2015; Kenna et al., 2016). Many of these genes have subsequently been found to carry ALS causal mutations, including a study by our own laboratory identifying a novel *TBK1* mutation (Williams et al., 2015). Further, the novel ALS gene *CCNF*, discovered by our research team, was also found to have a genetic burden in SALS compared with controls (Williams et al., 2016b). This demonstrates the potential for genetic burden to contribute to the cause of ALS, and to implicate novel genes in the cause of ALS. As the availability of NGS datasets from ALS patients grows, such genetic burden-based studies are likely to facilitate the discovery of more novel ALS genes.

8.2.1.4 Structural variation

Structural variants (SVs) are genetic alterations affecting chromosome structure and involve large sections of DNA. Various types of SVs have been reported, and include inversions, translocations and genomic imbalances, namely large insertions or deletions, as well as copy number variants (CNVs) and repeat expansions. SVs can arise from errors during cell division or incorrect DNA repair at any point of the cell cycle, including in post-zygotic cells, and often occur in repetitive and duplicated genomic regions (Weckselblatt and Rudd, 2015). The consequences of SVs can vary greatly depending on their genomic location, with many thought to play a regulatory role in gene expression, while any SV that disrupts a gene severely compromises transcription (Weckselblatt and Rudd, 2015). SVs have been associated with numerous developmental disorders (Weckselblatt and Rudd, 2015) and have also

been implicated as a cause or phenotypic modifier of a number of neurodegenerative diseases (reviewed in Lee and Lupski, 2006). Repeat expansions are a known cause of neurodegenerative conditions including several spinocerebellar ataxias (Banfi et al., 1994; Orr et al., 1993; Pulst et al., 1996), Kennedy's disease (La Spada et al., 1991) and Huntington's disease (1993). Importantly, the most common known cause of ALS, expansion of *C9orf72*, is a repeat expansion, a type of CNV. While full expansion of a repeat expansion in *ATXN2* causes spinocerebellar ataxia type 2 (Pulst et al., 1996), intermediate lengths of this repeat expansion have also been associated with increased risk of developing ALS (Elden et al., 2010). As such, there is compelling evidence to further investigate the contribution of novel repeat expansions or CNVs to ALS. Unrecognised repeat expansions or CNVs may underlie the cause of ALS in a proportion of the 40% of FALS with unknown mutations, and may also contribute to disease onset or progression in SALS patients. Further, other structural variants such as translocations or inversions, have also been implicated in diseases such as Duchenne muscular dystrophy (Oshima et al., 2009), and may also contribute to ALS.

8.2.1.5 Epigenetics

Epigenetic mechanisms, which act to regulate gene expression, include DNA methylation, histone modifications and chromatin remodelling. They exert their influence on processes such as DNA replication and repair, as well as RNA transcription and chromatin formation, which ultimately control downstream protein translation (Belzil et al., 2016). Epigenetic signatures are dynamic and change over time (Feil and Fraga, 2011; Handel et al., 2010). As such, an accumulation of epigenetic alterations over a lifetime fits well with the late and variable age of onset of ALS (Belzil et al., 2016). Epigenetic patterns have also been shown to be altered in response to aging and various environmental exposures (reviewed in Feil and Fraga, 2011). As noted in Chapter 1, Section 1.3.2, many environmental factors have been suggested to increase the risk of ALS. Therefore epigenetic alterations may provide the crucial link between environmental exposure and the onset of disease, or the phenotypic heterogeneity observed between ALS patients (including relatives with identical causal mutations) (Belzil et al., 2016). To fully elucidate the causes of ALS (particularly SALS), both genetic and environmental components, and therefore epigenetic modification, will need to be considered.

Though still a relatively new area of research, a number of studies support an epigenetic contribution to ALS. To-date, DNA methylation is the best understood

and therefore most researched epigenetic mechanism. DNA methylation is upregulated in the promoter region of pathogenic *C9orf72* expansion alleles (Belzil et al., 2014; Xi et al., 2013), as well as the expanded hexanucleotide repeat region itself (Xi et al., 2015). Comparisons between neurological tissue from SALS patients and neurologically normal individuals have also identified differentially methylated genomic regions (Figueroa-Romero et al., 2012; Morahan et al., 2009) and global DNA methylation increases in SALS patients (Chestnut et al., 2011; Figueroa-Romero et al., 2012). Recently, our own laboratory identified epigenetic changes between disease-discordant MZ co-twins (Manuscript A4, Appendix A.5.4). Most compelling was the identification of accelerated epigenetic aging in the ALS affected co-twin, which has also reported by Young et al. (2017). Together, these findings support a role for epigenetic modification in regulating the onset and/or progression of ALS.

8.2.1.6 The multi-step hypothesis of ALS

It is generally accepted that ALS is a complex disease that results from the interplay between genetic and environmental factors. Recently, it was postulated that ALS is a multi-step process in which several sequential events (or steps) are required to trigger the onset of disease. An accumulation of effects that cause disease is supported by the late and highly variable age of onset in ALS patients. The statistical model for the multi-step process is based on a linear relationship between the log values for the age of onset and incidence of disease (Armitage and Doll, 1954). Initially, such a relationship was reported by Al-Chalabi et al. (2014) among 6,274 ALS patients (both FALS and SALS cases) which suggested that an average of six steps were required to initiate the onset of disease. Subsequently, the hypothesis that ALS causal mutations account for multiple steps was explored by Chio et al. (2018), which is strongly suggested by the lower age of onset observed in FALS patients compared with SALS. This analysis showed that *SOD1* mutations effectively accounted for four of the six steps, while pathogenic expansion of *C9orf72* and *TARDBP* mutations contributed three and two steps respectively.

This also supports the idea that common ALS genes such as *SOD1*, *C9orf72* and *TARDBP* elicit a large effect on the mechanism of disease, but gene mutations in families with reduced penetrance are likely to have smaller effects, and account for fewer steps in this multi-step process. This was also reflected by data from the *in silico* pipeline for assessment of potential pathogenicity, where highly penetrant *SOD1* mutations scored higher than less penetrant ALS mutations such as *UBQLN2* and

CCNF. The multi-step process hypothesis is also consistent with an oligogenic basis of disease, where multiple ALS gene mutations may each account for more than one step of the multi-step process.

The remaining steps of this model are likely to be environmental exposures, many of which have previously been suggested as associated with ALS (See Chapter 1, Section 1.3.2). Given that genetic predisposition to ALS appears to account for multiple steps, and fewer environmental exposures will be needed to trigger disease onset in mutation-carrying patients, future studies should investigate environmental risk factors in genetically predisposed ALS patients (Chio et al., 2018). This can also be extended to the investigation of epigenetic contributions to disease, as was discussed above in Section 8.2.1.5.

8.3 Next-generation sequencing

NGS data were used for all genetic investigations conducted in this project. Such datasets are an asset for genetic research, and the utility of NGS data has been demonstrated throughout this project. NGS data have facilitated investigations of known and candidate ALS genes, as well as novel gene discovery in ALS families and disease-discordant MZ twins. However, these analyses have also highlighted many of the pitfalls of NGS data, and the necessary considerations for accurate NGS analysis. Chapter 3 highlighted the bioinformatic barriers to effective NGS analysis, and presented numerous custom strategies that were developed to overcome such barriers. While the benefits of NGS-based approaches to genetic discovery far outweigh the drawbacks, it is vital that scrupulous validation, interpretation and caution are applied when working with these datasets. The following sections will discuss vital points that must be considered in NGS analyses that have been highlighted throughout this thesis.

8.3.1 WES vs WGS

Many factors should be taken into account when deciding to choose whole-exome or whole-genome sequencing. These include cost, the purpose of sequencing, data quality and bioinformatics burden. When NGS first became available, sequencing costs were extremely high and WGS was prohibitively expensive. Today, the price gap

between WES and WGS sequencing continues to narrow, and WGS is expected to eventually become more economical by avoiding the cost of targeted exome capture. However, the costs associated with computing performance are not declining in-line with sequencing costs, thus limiting the accessibility of WGS (Laurie et al., 2016).

As part of this thesis, both WES and WGS were utilised for genetic analysis of ALS patient cohorts. These comparative analyses highlighted the vast increase in genetic sequence information produced by WGS compared with WES. WGS identifies at least ten-fold more variants than WES (Gilissen et al., 2014). This increase in data volume has implications for both the storage and analysis of genetic data. In this project, WES data was able to be stored on standard storage devices, however WGS data required specialised storage solutions using large shared-memory systems. Similarly, while bioinformatic scripting strategies for WES were readily performed using standard desktop computing systems, this was not possible when analysing WGS data, and high-performance computing systems were required. This was also reflected by the process of ANNOVAR annotation, which was seamlessly performed for WES data, but required the development of a custom scripting strategy for WGS data (Chapter 3, Section 3.5.1). Further, the exome has far fewer repetitive sequences and is more highly conserved compared with non-coding regions (Meynert et al., 2014), meaning that many more variants are identified per kilobase of DNA sequence in WGS than WES. Many repetitive and low complexity genomic regions targeted only by WGS are more prone to sequencing errors and the likelihood of false variant identification (discussed in detail below in Section 8.3.3). Therefore, analysis and interpretation of these regions is far more complicated than that of coding regions. Consequently, when deciding between WES and WGS, the value and purpose of sequencing non-coding regions must be considered.

While the exome accounts for less than ~2% of the genome, these coding regions harbour 85% of known disease causing mutations (Liu et al., 2015). As such, WES is a popular, cost effective choice as a first-line approach to identify disease mutations (Bamshad et al., 2011), with the added advantage of avoiding complex bioinformatics analyses. Indeed, as part of this project, WES was used for novel ALS gene discovery in FALS. However, the abundance of known disease mutations in the exome may simply reflect that it has been far more extensively studied than other genomic regions.

WGS has greater uniformity of coverage across the genome than WES has across the exome (Belkadi et al., 2015; Lelieveld et al., 2015; Meynert et al., 2014), which

contributes to the increased SNP sensitivity of WGS (Meynert et al., 2013). WGS also obtains better uniform coverage in GC-rich regions and this is increased further with PCR-free WGS (Meienberg et al., 2016). Since exome-capture probes are complementary for reference alleles, wild-type sequences may be overrepresented in raw read data from WES, and potentially mask heterozygous SNPs (Meynert et al., 2013). This bias dictates that approximately three-fold more read depth is required to obtain accurate genotyping data from WES than WGS (Lelieveld et al., 2015; Meynert et al., 2014). Uniform coverage also allows WGS to detect more complex genetic aberrations such as structural variation (including copy number variation).

Reports have shown that 10-19% of the exome may not be adequately covered by WES (Liu et al., 2015), and the false positive rate in coding regions is also higher for WES than WGS (Belkadi et al., 2015). WGS and WES do show high concordance (98.03-99.46%) for genotypes at polymorphic SNPs, with discordant sites generally reported as heterozygous by WGS and homozygous by WES (Belkadi et al., 2015; Laurie et al., 2016; Meynert et al., 2014). However, concordance between indel calls is much lower at 65.76%-84.85% (Laurie et al., 2016). Interestingly, the total proportion of false positive indels called in WES and WGS data is similar, at 44% and 46% respectively (Belkadi et al., 2015). This reflects the difficulties in detecting indels using short read NGS sequencing, and the need for improved strategies for indel detection and calling.

Another important consideration when choosing between WES and WGS is variant interpretation. Coding variation is well characterised, and clear cause and effect relationships can often be established. Therefore, biologically meaningful conclusions are more readily reached when studying coding regions. Conversely, little is known about the non-coding regions of the genome. Intergenic regions, which contribute to the majority of genomic sequences are poorly understood, and were once referred to as "junk DNA". While these regions are known to contain regulatory elements which do contribute to function and phenotype (as described in Section 8.3.4), any mutations identified here are less likely to be linked to informative biological functions, and are of little utility until bioinformatic tools and cell-based assays are developed to better understand their contribution to disease. The interpretation of non-coding variation will be further discussed in Section 8.3.4.

8.3.2 Bioinformatics processing

Bioinformatic analysis is arguably the most important component of an NGS workflow. In order to gain meaningful insights from NGS data, robust bioinformatic analysis strategies are required. These include the standard processing steps of sequence alignment, variant calling and annotation, as well as project-specific genetic analysis and filtering. Various analysis strategies have been developed and presented throughout this thesis. Most notably, in Chapter 3, numerous scripting strategies were developed to address common requirements for NGS data analysis, while more specialised scripting strategies were developed as part of Chapters 4 – 7 to perform project-specific genetic analyses.

Processing of raw NGS data is now largely standardised, and includes quality control, read alignment and variant calling. Quality control of NGS data is an integral step of the NGS workflow, to ensure reliable downstream results (Pabinger et al., 2014). Base calling errors, poorly defined indels, poor quality reads and adapter contamination all hinder sequencing quality and reliability (Dai et al., 2010). NGS platforms are also susceptible to instrument and chemistry failures, and such mishaps are not uncommon (Pabinger et al., 2014). Trimming and removal of low quality sequencing reads is based on a number of key metrics including base quality scores, nucleotide distributions, Kmer length, N content and GC bias (Cox et al., 2010; Pabinger et al., 2014).

Various data processing software tools are available for read alignment and variant calling, to transform raw NGS read data into meaningful genetic variant data. Each of these tools has its own strengths and weaknesses based on parameters such as error rate, alignment speed, memory, sensitivity and accuracy (Li et al., 2015). As part of this project, the Burrows Wheeler Aligner (BWA) (Li and Durbin, 2009, 2010) and the Genome Analysis ToolKit (GATK) variant caller (McKenna et al., 2010) were employed. These tools are adopted as part of the GATK best practices (McKenna et al., 2010), which are generally regarded as the gold-standard processing pipeline, and have consistently been reported as top performing tools of choice (Hwang et al., 2015; Liu et al., 2013; Mielczarek and Szyda, 2016). In Chapter 7, NGS processing was also performed using the Isaac alignment and variant calling pipeline that is optimised for processing speed (Raczy et al., 2013). However, the Isaac pipeline has not been widely adopted, as evidenced by its absence from many studies comparing alignment and variant calling tools (Hwang et al., 2015; Laurie et al., 2016; Liu et al., 2013; Mielczarek and Szyda, 2016; Pirooznia et al., 2014; Yu and Sun, 2013). Studies

comparing various alignment and variant calling tools suggest that the BWA-GATK pipeline is favourable for routine application in NGS-based studies, though to obtain a high confidence variant call set, the results of various pipelines should be overlapped (Hwang et al., 2015; Liu et al., 2013; Mielczarek and Szyda, 2016; O'Rawe et al., 2013).

ANNOVAR (Wang et al., 2010) was chosen for annotation due to its ease of application and incorporation of information from a wide range of applicable databases. The RefSeq annotations provided by ANNOVAR are desirable for their compatibility with the tools available on the UCSC web browser (`https://genome.ucsc.edu/`), which was used extensively for genomic reference information as well as primer design. The integration of control database information from dbSNP, 1000Genomes, ExAC and gnomAD was also highly useful to facilitate variant filtering in Chapters 5 and 6. The ability of ANNOVAR to incorporate results from *in silico* protein prediction tools using the Database for Non-Synonymous Snps' Functional Predictions (dbNSFP, Liu et al., 2011) was also valuable, as these predictions were utilised as part of the *in silico* pipeline developed in Chapter 6 to assess the potential ALS pathogenicity of candidate mutations. Additionally, the tab-delimited output format was also a useful feature to increase the ease of variant visualisation and filtering.

8.3.3 Sequencing artefacts

Despite an accuracy rate of more than 99.9%, the massively high-throughput nature of NGS inevitably produces thousands of sequencing errors (Fernandez-Marmiesse et al., 2018; Lohmann and Klein, 2014). With the increasingly widespread use of NGS technologies in genetic discovery as well as in the clinic for genetic diagnosis, understanding the sources of sequencing artefacts is vitally important. While NGS is generally perceived as a single technology, it must be remembered that it consists of three distinct modules including library preparation, sequencing and bioinformatic processing (Daber et al., 2013). Each of these modules is prone to error, and therefore, sequencing artefacts can be introduced at any stage of the NGS pipeline. Sequencing artefacts can be false positive (an introduced, incorrect sequence variant) or false negative (a real variant that has not been detected) variant calls. To validate variants identified by either WES or WGS, Sanger sequencing was performed for each candidate mutation identified in this project using patient DNA samples, PCR amplification and Sanger sequencing. This validation revealed that false positive sequencing artefacts were abundant across the NGS datasets utilised throughout this thesis. While the

exact source of many of these artefacts cannot be determined without replication of NGS data generation, it is expected that a large proportion of artefacts resulted from bioinformatics processing, as was seen for the discordant variants identified between co-twins/triplets in WGS data (Paper III, Chapter 7). Unfortunately, Sanger sequencing validation is only able to detect false positive variants in NGS data, as false negative variants are never identified and thus never validated (Daber et al., 2013). Section 8.3.3 will provide an overview of the potential sources of error leading to the introduction of sequencing artefacts.

Throughout this thesis, more sequencing artefacts were identified in WGS data compared to WES data. In Chapters 5 and 6, Sanger sequencing of candidate gene variants showed that 20 of 85 variants (~23.5%) from WES data, and 13 of 16 variants (~81.25%) from WGS data, were false positives. Additionally, in Chapter 7, tens of thousands of discordant variants were identified between ALS-discordant MZ twin pairs using WGS data, all of which were concluded to be sequencing artefacts following extensive validation. The increased abundance of sequencing artefacts in WGS data compared with WES data can be attributed to a combination of factors. These include the increased scale of sequencing, the low complexity of many non-coding genomic regions, and lower average sequencing coverage of the WGS data (30X) in contrast to WES data (100X). Each of these factors will be discussed in further detail below in Section 8.3.3.

Interestingly, when WES data from FALSmq28 is removed from the above figures, the number of WES candidate mutations found to be false positives was reduced to just five of 85 (~5.88%). The increased number of false positive variants in the WES data from FALSmq28 is likely due to the fact that the raw WES data for the three members of this family underwent single-sample variant calling, while all other WES data (n=137) underwent joint variant calling. Multi-sample calling has been shown to improve both the sensitivity and accuracy of variant calls (Liu et al., 2013), therefore the FALSmq28 WES false positive variants are likely attributable to a reduction in these metrics. A proportion of the sequencing artefacts in FALSmq28 WES may also be attributable to errors during library preparation or sequencing (as discussed below), as this sequencing data was also generated separately to any other dataset.

Up to 90% of the genome can be confidently genotyped (Laurie et al., 2016; Telenti et al., 2016; Weisenfeld et al., 2014; Zook et al., 2014), with a false discovery rate of just 0.0008 for variant identifications (Telenti et al., 2016). However, the remaining

10% is almost impossible to accurately genotype using current technology (Laurie et al., 2016; Telenti et al., 2016; Weisenfeld et al., 2014; Zook et al., 2014). This portion of the genome consists of low complexity, repetitive or duplicated genomic regions (such as paralogues, pseudogenes, transposable elements, tandem repeats, segmental duplications and complex rearrangements), or regions with large structural variants (Li, 2014; Linderman et al., 2014; Weisenfeld et al., 2014; Yu and Sun, 2013; Zheng and Grice, 2016). Sequencing artefacts arising in these regions are attributable to a number of factors, most notably amplification and sequencing issues caused by high GC-content, as well as difficulties encountered by alignment and variant calling tools. The following sections provide further details of the sources of NGS errors, and though these factors are most relevant for these difficult to sequence regions, they also cause sequencing errors in other genomic regions, albeit at a much lower rate.

8.3.3.1 Sources of sequencing artefacts

Library preparation

Whole-genome amplification (WGA) applies PCR to a genomic DNA sample for which an insufficient quantity of DNA is available. In Chapter 5, a novel candidate gene variant in *DAGLB* was identified in WES data generated from a patient DNA sample that had previously undergone WGA. Fortunately, this patient also had a small amount of non-amplified DNA available. Sanger sequencing showed that while this candidate mutation was present in the WGA DNA sample, it was absent from the non-amplified DNA sample. Therefore, it was deemed a PCR artefact. Further, this candidate mutation showed a high potential for ALS pathogenicity using the *in silico* pipeline. This result demonstrates the potential for WGA, and PCR in general, to introduce errors to NGS data, and highlights that such errors have the potential to exhibit characteristics suggestive of a pathogenic nature. Therefore, any NGS data generated from WGA samples must be treated with caution.

The introduction of mutations during WGA are similar to PCR errors. Indeed, PCR amplification prior to NGS sequencing has been shown to drastically increase sequencing error rates by introducing sequencing template base errors (Li, 2014). This was indeed the case in Chapter 7, where the one twin set that underwent PCR amplification during WGS library preparation showed up to 10-fold more discordant variants (later concluded to be sequencing artefacts) than any of the three twin/triplet sets that underwent PCR-free WGS. PCR base incorporation errors may be polymerase-derived errors, particularly those in poly-A or poly-T runs (Brockman

et al., 2008; Li, 2014). Sequencing template bias in amplified libraries may also be introduced by PCR, due to annealing difficulties in GC-rich regions (Aird et al., 2011), or higher affinity for particular alleles at polymorphic sites, which in turn affect downstream variant calls (Yu and Sun, 2013). PCR amplification prior to NGS is however necessary to increase the availability of sequencing templates in some circumstances, particularly for WES after exome capture where just ~2% of the genome is captured, leaving very small quantities DNA available to act as sequencing templates.

In addition to PCR artefacts, library preparation is error prone due to the creation of chimeric templates from adapter sequences that have not been properly cleaved, adapter/primer dimers or biases, 3' capture bias or the inclusion of adapter sequences in sequencing templates (Kircher et al., 2011; Robasky et al., 2014; Yu and Sun, 2013). Such errors may create contaminant or incorrect sequences, or introduce coverage biases, which may result in the misrepresentation of the true genetic sequence.

Sequencing

During the sequencing phase of NGS, multiple factors may cause incorrect base incorporation or base calling, leading to flawed sequencing reads. Sufficient sequencing cluster intensity is vital for correct base calling, though this intensity can be reduced by inadequate extension product growth during bridge amplification, inefficient hybridisation of sequencing primers, or degraded fluorophores (Kircher et al., 2011; Yu and Sun, 2013). Signal intensity may also be diminished, or incorrect, if sequence read synthesis from the individual template copies belonging to the same sequencing cluster becomes de-synchronised (Nielsen et al., 2011). Such de-synchronisation is amplified in each subsequent sequencing cycle, and therefore base calling becomes less accurate in later cycles (Nielsen et al., 2011). Contaminant sequence reads may result if molecules such as chemistry crystals, dust or lint particles are recognised as sequencing clusters by the sequencing instrument (Kircher et al., 2011). Such sequencing read errors have downstream complications, causing misalignment and incorrect variant calls during bioinformatics processing (Robasky et al., 2014). Low complexity regions suffer from the added difficulties that their repetitive nature and high GC-content cause sequencing fidelity errors, and are therefore generally sequenced at very low read depths (Meynert et al., 2014).

Bioinformatics processing

Errors arising from the bioinformatics processing of raw NGS sequencing reads were investigated in Chapter 7. Extensive validation efforts, together with the application

of multiple processing pipelines, suggested that the discordant variants identified between co-twins/triplets were most likely artefacts of bioinformatics processing. These findings reflect the high level of discordance in variant calls between different alignment and variant calling tools (Bao et al., 2011; Hwang et al., 2015; Laurie et al., 2016; Li, 2014; Liu et al., 2013; Mielczarek and Szyda, 2016; O'Rawe et al., 2013; Tian et al., 2016; Yu and Sun, 2013). Therefore, sequencing artefacts arising from the bioinformatics processing of NGS data are a common feature of NGS datasets.

The majority of errors introduced by bioinformatics processing tools are found in the 10% of the genome that is notoriously difficult to genotype (GC-rich sequences and repetitive and duplicated elements) (Laurie et al., 2016; Weisenfeld et al., 2014). Alignment is particularly difficult in these regions, as due to their short length, alignment tools may incorrectly map a sequence read to two or more reference genome locations that have highly similar sequence identity (Li, 2014; Linderman et al., 2014; Weisenfeld et al., 2014; Zheng and Grice, 2016). Such reads are either discarded or aligned to the "best match" genomic region. This results in regions with insufficient read coverage and/or dubious alignments, both of which lead to incorrect variant calls (Fernandez-Marmiesse et al., 2018; Laurie et al., 2016; Li, 2014; Linderman et al., 2014; Weisenfeld et al., 2014; Zheng and Grice, 2016). Further, alignment is also difficult in genomic regions where the individual's sequence deviates from the reference genome, such as those regions with many SNPs in close proximity (Nielsen et al., 2011; Reinert et al., 2015; Tian et al., 2016). Additionally, variant callers have also been shown to be better at calling dbSNP variants than novel variants (O'Rawe et al., 2013; Yu and Sun, 2013). Both alignment and variant calling tools are far more reliable for the remaining 90% of the genome, with concordance between different tools for these regions in the vicinity of 90% for SNPs and 60% for indels (Hwang et al., 2015; Laurie et al., 2016; Linderman et al., 2014; Popitsch et al., 2017).

Read depth

Read depth is a critical factor in NGS, as higher coverage leads to more accurate variant calling, and low coverage data can lead to incorrect variant identification (Linderman et al., 2014). It is also important to note that coverage is not uniform across the genome, with some regions being sequenced more readily than others, so that a sequencing coverage of 30X indicates the average number of reads mapped to any position across the genome, rather than the minimum number of reads mapped to every position. A 30X sequencing coverage, as used here for WGS, equates to most sites being covered by at least 10 reads (Lohmann and Klein, 2014), however some

sites may not be covered by any reads. This non-uniformity extends to allele coverage, so that one allele may be sequenced more times than the other, reflecting a false allelic distribution, which may result in incorrect variant calls (Lohmann and Klein, 2014). Estimates suggest that a minimum coverage depth of 33-50X (i.e. minimum coverage at each site, not average coverage across the genome) is required to detect all SNPs and small indels using WGS (Ajay et al., 2011; Bentley et al., 2008), however due to the difficulties associated with identifying larger indels, a much higher coverage will be necessary to identify these in NGS data.

8.3.3.2 Sequencing artefacts and Sanger sequencing validation

The majority of NGS variants that were found to be sequencing artefacts after Sanger sequencing in this project were located in low complexity, repetitive or duplicated genomic regions. In addition to impacting NGS accuracy, these characteristics also complicate primer design for PCR. Over 15,000 pseudogenes have been been identified in the human genome (Cunningham et al., 2015). Therefore, primers targeting these elements are likely to amplify numerous regions across the genome. This was found to be the case when designing primers for candidate mutations falling in low complexity, repetitive or duplicated regions. To combat this difficulty, approaches to increase specificity were employed. These included designing primer pairs to incorporate SNP variants, as well as using touch-down thermocycling conditions and nested PCR. In many cases, nested PCRs were required to obtain a single PCR product. This approach first involved the amplification of a 700-1000bp genomic region flanking the target variant. This long PCR product was then used as a template for a subsequent reaction targeting a smaller, more specific region of interest, which would otherwise amplify numerous loci from a genomic DNA sample. Even when such approaches were utilised, many PCR products targeting indel variants produced poor quality Sanger sequencing data. In these cases, a fragment length analysis approach was employed to resolve the size of the product, to confirm the presence or absence of the indel variant. All variants directly sequenced using nested PCRs or fragment length analysis were found to be false positive variants, demonstrating the high error rates of NGS in low complexity genomic regions. Further, Sanger sequencing showed that all 24 indel candidate mutations from FALSmq28 were false positive variants, reiterating the difficulties encountered by NGS in accurately identifying indel variants.

8.3.3.3 Additional considerations for dealing with sequencing artefacts

In order to avoid sequencing artefacts, some studies have suggested that filters should be applied to remove the 10% of the genome that cannot be reliably genotyped, or more simply, remove genomic regions of low complexity (Popitsch et al., 2017). However, such filtering approaches may also discard important variant calls. Other studies have instead suggested that overlapping the variant call sets from multiple variant calling tools will increase specificity, and produce a high confidence set of variants (Li, 2014; Liu et al., 2013; Popitsch et al., 2017). In Chapter 7, these approaches were combined to conclude that no reliable discordant variants had been identified between ALS-discordant co-twins/triplets.

Regardless of whether a region is difficult to sequence or not, single nucleotide variant (SNV) calls are more reliable than those for indels (Laurie et al., 2016; O'Rawe et al., 2013). Indels are notoriously difficult to identify using NGS, with reports showing that only ~50% of indels identified by WES and WGS can be validated by Sanger sequencing (Belkadi et al., 2015). Indeed, many of the NGS indel variants identified throughout this project were not validated by Sanger sequencing (as described above in Section 8.3.3.2), or the position of the NGS indel variant was called inconsistently. The difficulty in aligning sequencing reads containing indels to the reference genome means that indels are often mapped incorrectly (Daber et al., 2013; Li, 2014; Pabinger et al., 2014). Sequencing reads that contain large indels are frequently discarded from bioinformatics processing, leading to false negative variant calls (Daber et al., 2013; Pabinger et al., 2014). Indels are also more susceptible to replication errors than SNVs (Li, 2014).

False negatives are not detectable without repeating NGS. It is possible that a mutation that caused ALS may not have been detected with the NGS strategy used in this project. This may be the case in FALSmq28 in Chapter 6, where no candidate mutations remained after family-based analysis and Sanger sequencing validation, or in the twin pairs described in Chapter 7 where no somatic *de novo* mutations were identified between ALS-discordant co-twins/triplets.

8.3.4 Assessment of variant pathogenicity

An unprecedented amount of genetic variation has been uncovered using NGS technologies, and this can often lead to many candidate mutations from gene discovery

projects, as shown in this thesis. Interpreting the biological relevance of variants is now a major bottleneck in disease gene discovery (Lohmann and Klein, 2014). Many variants identified by NGS may be predicted to play a role in disease, a problem that has been labelled the "narrative potential" of the human genome (Goldstein et al., 2013). For example, after sequencing 10,000 individuals, Telenti et al. (2016) showed that an average of more than 8,000 novel variants were present in single human genome. Many interactions may be identified that link novel variants, or affected proteins, with known disease-linked proteins or mechanisms, though whether these truly contribute to disease pathogenesis cannot be established without appropriate *in vitro* and *in vivo* modelling. The scale of this task is typically beyond the scope of current functional assays, particularly when considering non-coding variants.

While NGS approaches have facilitated many gene discoveries, there have been instances where the claimed pathogenicity of genetic variants has been called into question. This reflects the difficulty in assessing and interpreting the pathogenicity of variants identified by NGS. In fact, systemic reviews suggest that up to 27% of variants reported as disease-linked mutations have later been revealed to be benign polymorphisms present in population controls, or to have insufficient evidence to be labelled as pathogenic (Xue et al., 2012). Some family studies, and studies of proband or sporadic patient cohorts have labelled variants as pathogenic based solely on the observation that they fall within a known disease gene or disease related gene. A prime example is *SOD1*, where over 180 putative pathogenic mutations have been reported across the 462 nucleotides of the coding sequence, many of which have been reported in a single case without supportive segregation or replication data. As genetic discoveries are translated to the clinic to improve diagnostics and treatment decisions for patients (Quintans et al., 2014; Richards et al., 2015), inaccurate assignment of pathogenicity can have severe consequences for patients, by initiating incorrect prognostic, therapeutic or reproductive advice (MacArthur et al., 2014). Therefore, accurate assignment of pathogenicity is crucial for effective translation of genetic discoveries into the clinic. From a research perspective, resources may also be wasted by supporting research programs that are based on false assignments of pathogenicity, with misallocation of resources for studies targeting genetic variants, or proteins, that lack adequate evidence for their role in disease (MacArthur et al., 2014).

Clear guidelines are required for assessing variant pathogenicity and prioritisation. MacArthur et al. (2014) and Richards et al. (2015) separately reviewed this from a research and clinical perspective, respectively. Both included discussions of the

evidence required from genetic and functional sources to implicate a variant as being a pathogenic mutation (MacArthur et al., 2014; Quintans et al., 2014; Richards et al., 2015). Foremost in family-based studies is segregation of a candidate mutation with disease within a family (where possible), which may be complemented by the presence of identical or additional novel mutations within the same gene in other families or patient cohorts. Absence of the variant from large numbers of population controls is also necessary for genetic confidence of variant pathogenicity (MacArthur et al., 2014; Richards et al., 2015). Functional characteristics of both the gene and variant should also support the pathogenicity of a variant (MacArthur et al., 2014; Richards et al., 2015). This functional data can be gathered using *in silico*, *in vitro* and *in vivo* approaches. As part of this thesis, the genetic evidence available to implicate a variant in disease has been extensively scrutinised. Additionally, an *in silico* pipeline was developed to assess the potential pathogenicity of candidate mutations based on functional characteristics, in order to prioritise those most suited for functional assessment using *in vitro* and *in vivo* models.

8.3.4.1 *In silico* pipeline for assessment of variant pathogenicity

The pipeline presented in Chapter 6 incorporated a range of gene and variant level functional characteristics to prioritise candidate mutations within each ALS family. Analysis of amino acid conservation across species and prediction of damaging protein effects have been reported as vital tools to assess variant pathogenicity (MacArthur et al., 2014; Richards et al., 2015). High evolutionary conservation suggests the integral importance of an amino acid residue, implying that divergence would lead to detrimental changes in protein function. Protein prediction programs generally utilise sequence homology and/or protein structure information to predict the likelihood that variant is damaging (Frousios et al., 2013) and thereby potentially pathogenic (MacArthur et al., 2014; Quintans et al., 2014). Gene expression levels in the disease affected tissue may also implicate genes that are more likely to be relevant to pathology (MacArthur et al., 2014). Finally, the natural abundance of variation within a gene has been reported as a good indicator of how well a gene will tolerate protein-altering variation without significant detrimental functional effects (MacArthur et al., 2014).

It has been argued that data from multiple *in silico* tools that assess the same characteristic can be interpreted as a single piece of evidence (MacArthur et al., 2014; Richards et al., 2015). While various tools that assess the same characteristic have substantial underlying similarities, each utilises a unique algorithm and set of

parameters to calculate their result (MacArthur et al., 2014; Pabinger et al., 2014; Richards et al., 2015). As such, the outputs from these tools are not always in agreement, and some tools perform better for particular types of variants or genes (Richards et al., 2015). Multiple tools and/or databases were used to assess each of the four characteristics incorporated as part of the *in silico* pipeline presented here, in order to ensure that a general representation of each characteristic was reported, and to avoid biases introduced by any individual tool.

The potential pathogenicity of 110 coding candidate mutations (identified by genetic analysis across the four small families analysed in Chapter 6) was assessed using the *in silico* pipeline developed here. This prioritised candidate mutations and determined which were most suited for downstream *in vitro* functional analysis. While the *in silico* approach cannot definitively prove the causality of a gene mutation, the findings of these analyses can be a great asset in determining the most likely cause and mechanism of disease within a family, and can be informative for other researchers who have made similar observations.

8.3.4.2 Non-coding variant interpretation

Non-synonymous variants can be assessed to predict downstream effects on protein structure and function, and potential pathogenicity, using the numerous *in silico* tools as described above (MacArthur et al., 2014). In contrast, non-coding variants identified by WGS present a new challenge as our understanding of their functional consequence is in its infancy. As non-coding regions account for more than 98% of the genome, millions of non-coding variants are present within each individual's genome. This was demonstrated in family FALSmq28, where over 10 million non-coding variants were identified by WGS for the three sequenced family members.

Many non-coding variants have regulatory roles that affect gene splicing and expression, yet the knowledge available in this space is still quite limited. Promoter variants may impact transcription, enhancer variants may affect transcription factor binding motifs, intronic and untranslated region variants can affect messenger RNA by altering stability or splicing patterns, and other non-coding variants may alter various classes of RNA molecules including long non-coding RNAs, microRNAs and small nucleolar RNAs (Zhu et al., 2017).

A number of resources for interpreting non-coding variation are beginning

to emerge. These include databases such as the Encyclopedia of DNA Elements (ENCODE) (Dunham et al., 2012) and the Roadmap Epigenomics Consortium (Kundaje et al., 2015), both of which provide data to assess the potential contribution of a genomic region to regulatory processes. Other prediction tools for non-coding variants including CADD (The Combined Annotation Dependent Depletion; http://cadd.gs.washington.edu/; Kircher et al., 2014), DANN (The Deleterious Annotation of genetic variants using Neural Networks tool Quang et al., 2015), GWAVA (The Genome-wide annotation of variants; https://www.sanger.ac.uk/sanger/StatGen_Gwava; Ritchie et al., 2014), FATHMM-MKL (Functional Analysis through Hidden Markov Models; http://fathmm.biocompute.org.uk/; Rogers et al., 2018; Shihab et al., 2015) and LINSIGHT (linear INSIGHT; https://github.com/CshlSiepelLab/LINSIGHT; Huang et al., 2017) are also available.

8.4 Important considerations for disease gene discovery using NGS datasets

As established above, while NGS provides an exciting opportunity to identify novel genetic factors underlying disease, there is an alarming potential for incorrect conclusions to be drawn from such studies. This may arise from the incorrect removal of disease causal variants during variant filtering, or incorrect assignment of pathogenicity to benign variants. Complex genetic diseases continue to pose challenges for genetic study design. Further, while large scale databases represent rich resources to aid genetic analyses, they too must be treated with caution. As such, various points must considered when evaluating the results of NGS-based studies, as discussed in the following sections.

8.4.1 Limitations of family-based analysis

The development of formal genome-wide statistical models may be required to differentiate between variants causing disease and those implicated by chance (MacArthur et al., 2014). Such models need to account for disease prevalence, disease penetrance, natural genic variation and mutation rates (MacArthur et al., 2014). While such models have been developed for identifying disease-associated variants in GWAS, no comparable models are yet available for monogenic diseases (MacArthur et al., 2014). For many diseases, including ALS, the patient sample sizes necessary to reach

statistical significance will require large international collaboration (MacArthur et al., 2014). However, for extremely rare variants or "private" mutations limited to a single family, significance may never be attained (MacArthur et al., 2014).

Failure to call a pathogenic variant following NGS will also confound segregation analysis. If this "missed" pathogenic variant lies close to another rare variant, the latter may be mistaken as a potential causal variant rather than merely a linked allele (Fernandez-Marmiesse et al., 2018; MacArthur et al., 2014). As such, segregation alone may not be sufficient to establish pathogenicity (MacArthur et al., 2014). This is relevant for WES data, where many variants go undetected. To overcome this issue, comprehensive haplotyping of the region found to carry a candidate mutation in an extended pedigree would be necessary.

8.4.2 International databases of next-generation sequencing datasets

The generation of NGS sequencing data from thousands of individuals across the world presents an exciting opportunity to further our understanding of the genetic variation underlying human disease and phenotypic traits. Databases of NGS data from control cohorts provide an invaluable resource for filtering benign genetic variation. Over 150,000,000 SNVs across the genome have been reported in dbSNP, while over 10,000,000 coding variants can be found in the ExAC database. This staggering amount of genetic sequencing data provides great power for genetic analysis.

While the established control databases have limitations (as discussed in Chapter 6), they are highly informative for many modern genetic research applications. The control databases were integral components of genetic analysis presented in this thesis. Without these databases, control screening would have been too costly and time consuming to screen all candidate gene mutations and would have been limited to far fewer control individuals than are available in the online databases.

Caution must however always be applied when utilising aggregated control databases. Many of the variants reported in these databases have not been validated by Sanger sequencing, and therefore may be sequencing artefacts. The collaborative nature of many of these databases also means that individual DNA samples have been sequenced with a variety of sequencing platforms and processed using various

bioinformatics pipelines. As has been discussed above in Section 8.3.3, the three distinct modules of the NGS pipeline can each introduce a plethora of sequencing artefacts. Without uniform protocols applied across all samples within NGS databases, variant frequency biases associated with distinct protocols may exist within these databases, which may confound association analyses.

Additional methods may also be required to address population stratification of rare variants, as these show stronger geographical clustering compared with common variants (Mathieson and McVean, 2012; O'Connor et al., 2013). Indeed, as part of this thesis, the effect of population stratification was an important consideration in genetic analyses. In Chapter 4, it was demonstrated that the genetic landscape of Australian FALS and SALS is unique when compared to other primarily Caucasian-based populations. This was highlighted when the most common *SOD1* mutation in North America, p.A5V, was found to be absent from Australian FALS. Similarly, just ~3% of Australian SALS was attributable to the *C9orf72* expansion compared with 7-10% in other European based SALS populations. The results of SNP association testing in Chapters 4 and 5 showed discordance when using international control cohorts from ExAC and gnomAD, compared with Australian controls from DACC and MGRB cohorts. These SNP association results were also inconsistently replicated using the Project MiNE cohort of ALS patients and controls, which is primarily derived from European populations. Finally, in Chapter 6, variant filtering with Australian controls resulted in a substantial reduction in candidate mutations, even following filtering using large international cohorts. These findings highlight the importance of using controls with the same genetic background as the disease cohort.

8.5 Ongoing and future work

The ultimate goal is to identify novel genetic causes of ALS among Australian patients. The work presented in this thesis represents significant progress towards this goal. Known and candidate ALS gene analysis in Australian patients demonstrated the unique genetic landscape of Australian FALS, and further identified eight candidate gene mutations potentially causing ALS, and 17 potentially disease-associated candidate gene variants. Novel gene discovery efforts substantially narrowed the search for the ALS causal mutation in each of five Australian ALS families, with just a handful of strong candidate mutations present in each, and highlighted the potential for more complex genetic variants to underlie ALS in families and disease-discordant MZ

twins. Numerous bioinformatic scripting strategies were developed by the candidate to perform this genetic analysis of NGS data, which can also be applied to future datasets.

8.5.1 Familial ALS

The FALS candidate gene screening strategy that was developed throughout this thesis will continue to be applied to Australian FALS patients to identify novel genetic causes of ALS. Novel family-based gene discovery efforts will also continue. This will include performing updated filtering of population-based variants from the lists of candidate mutations for each family analysed in this thesis. Additionally, this will include family-based analysis of newly recruited families with multiple informative individuals available. Where a sufficient number of family members have DNA samples available, this analysis will incorporate genome-wide linkage analysis. For families where a compelling candidate ALS gene mutation is identified, international FALS cohorts will be interrogated for mutations in this candidate gene. Further, any newly reported ALS genes will be screened through the full Australian FALS cohort to establish their contribution to Australian FALS.

8.5.1.1 *In vitro* and *in vivo* assessment of pathogenicity

Given that the genetic power of the four small families has been exhausted, alternative strategies are necessary to implicate the pathogenic mutations. As part of the research program in our laboratory, high priority candidate mutations (as assessed using the *in silico* pipeline presented in this thesis) will be assessed for potential ALS pathogenicity using an established functional analysis pipeline. These analyses will characterise a number of ALS relevant phenotypes at the cellular level. For example, expression constructs can be generated for each candidate gene in parallel with the wild-type gene as a control construct. Cell lines transfected with these constructs will be used to determine the toxicity and cellular localisation of candidate mutant proteins. A cell death assay, analysed by flow cytometry, can assess the toxicity of the mutant gene and the cellular localisation determined by fluorescent visualisation of the proteins. Further, the candidate constructs can be co-transfected with TDP-43, the primary constituent of the disease hallmark protein aggregates. This would identify any potential pathological interactions between candidate mutant proteins and TDP-43 that may signify disease relevant pathology. Histopathological characterisation of candidate proteins in patient neuronal tissues can assess their

expression in motor neurons including any colocalisation with known ALS proteins and hallmark pathological features.

Candidate mutations with compelling support for pathogenicity from genetic, *in silico* and *in vitro* analyses can be modelled using *in vivo* strategies. For example, zebrafish and mice have been successfully used to develop ALS models (reviewed in Picher-Martel et al., 2016; Van Damme et al., 2017). Facilities for the development of these models are in place as part of the the multidisciplinary Macquarie University Centre for motor neuron disease research. Animal models can provide strong evidence to support, or refute, the role of a candidate mutation in motor neuron death. These models may also be useful in longer-term pre-clinical studies.

8.5.2 FALSmq28

A variety of research strategies may be considered to seek the identity of the ALS causal mutation in family FALSmq28. Moving forward, all family-based analysis will focus on non-excluded linkage regions. Given the plethora of sequencing artefacts identified by WGS throughout this thesis, WGS can be repeated to a greater depth or with a different sequencing platform. The re-sequenced WGS data can be compared to, or combined with the original WGS data, in order to remove any artefacts introduced by the library preparation or sequencing phases of WGS. Analysis can also commence to identify any repeat expansions, CNVs, or other SVs, that may cause disease using the strategy discussed in Section 8.5.4.

8.5.3 Sporadic ALS

As was established in section 1.4.2, there exists a significant amount of evidence implicating genetic variation in the cause of SALS (Al-Chalabi and Hardiman, 2013; Andersen and Al-Chalabi, 2011). Increasingly large SALS WGS datasets offer a unique opportunity to identify novel ALS genes of small to modest effect using association-based approaches. During the later stages of this candidature, WGS from 635 Australian SALS patients became available for analysis. Unfortunately, due to time constraints, genetic analysis of SALS in this thesis was limited to four candidate genes (*CHCHD2*, *CHCHD3*, *CHCHD6* and *TIA1*), and the known ALS gene *CHCHD10*. Future analyses may include a comprehensive ALS gene screen, similar to that carried out for FALS in Paper I (Chapter 4, Section 4.3.1), as well as

candidate ALS gene screening. Repeat expansions and CNVs will also be investigated as a cause of SALS as discussed below in Section 8.5.4.

8.5.3.1 Rare variant combinations and their association with disease-risk and variable phenotypes

The missing heritability in current SALS genetic studies, and the complex, heterogeneous nature of ALS, point to numerous potential genetic and/or environmental factors that contribute to development of SALS. A search can commence for combinations of rare SNPs that associate with 1) the presence or absence of ALS (i.e. disease risk), 2) disease duration and 3) age of onset, using WGS data and associated clinical information. This can be achieved by applying statistical software programs such as those based on Limitless Arity Multiple Testing (LAMP), a p-value correction technique for combinations of multiple markers (Terada et al., 2013). This approach may be extended beyond rare SNPs, to also investigate non-synonymous SNPs, rare non-synonymous SNPs and CNVs.

Disease-risk SNP combinations

A GWAS-based approach can be applied to identify combinations of rare SNPs associated with SALS, using the LAMPLINK software (Terada et al., 2016), which incorporates the LAMP methodology with the widely used GWAS software, PLINK. This approach requires a case-control analysis. Control WGS data from MGRB and Project MinE, each containing over 1000 control individuals, would be suitable for this analysis.

Prognostic SNP combinations

This analysis would use a LAMP approach optimised for survival analysis with log-rank testing (Relator et al., 2018), and involve a discovery and validation stage using SALS patient WGS data only. Patients with disease durations that place them in the top 10% of longest or shortest survivors in the cohort can be stratified, and WGS data from these two patient subsets used to identify combinations of SNPs that associate with slowly or rapidly progressive disease. These SNP combinations can then be interrogated in WGS data from the remaining 80% of patients, to assess whether disease duration may be predicted using this model. A similar strategy can also be applied to identify SNP combinations that influence the age of disease onset in SALS.

8.5.4 Copy number variation in ALS

Copy number variants (CNVs) are structural variants where the number of copies of a DNA segment varies. CNVs are a common feature of the human genome and a great source of genetic variation (Zhang et al., 2009). While the majority of known ALS mutations are nucleotide level coding mutations, the most common known cause of disease, the pathogenic expansion of an intronic hexanucleotide repeat *C9orf72*, is arepeat expansion, a type of CNV. Repeat expansions have also been identified as a cause of other neurodegenerative conditions including the spinocerebellar ataxias (Banfi et al., 1994; Orr et al., 1993; Pulst et al., 1996), Kennedy's disease (La Spada et al., 1991) and Huntington's disease (1993). Additionally, intermediate length repeat expansions in *ATXN2* have also been associated with increased disease-risk for ALS (Elden et al., 2010). Therefore, together with the recent difficulties of identifying novel nucleotide level mutations, this suggests that repeat expansions or other types of CNVs may play a pathogenic role in ALS. Unrecognised CNVs may underlie the cause of ALS in a proportion of the 40% of FALS with unknown mutations, and may also contribute to disease onset or progression in SALS patients. Further, other structural variants such as translocations or inversions, have also been implicated in diseases such as Duchenne muscular dystrophy (Oshima et al., 2009), and may also contribute to ALS.

CNVs have been understudied in ALS WGS studies, largely because of the limitations of existing software to identify CNVs on a genome-wide scale. Nevertheless, existing programs such as CNVnator (Abyzov et al., 2011) and Lumpy (Layer et al., 2014) could be used to seek CNVs in families, FALS probands, SALS patients and disease-discordant MZ twins. Each of these programs utilises different properties of raw WGS read data to determine the presence of a CNV. The MetaSV program (Mohiyuddin et al., 2015) can be used to overlap the CNVs identified by each program, to create a set of higher confidence CNV calls for each individual. The exome hidden Markov model (XHHM) program (Fromer et al., 2012) offers a means of CNV analysis with WES data, although the identification of CNVs from WES is less reliable than that from WGS.

In ALS families, including FALSmq28, shared variant analysis could be applied to the CNV data to identify those CNVs that segregate with disease. Where applicable, this analysis would be limited to non-excluded linkage regions. The CNVs identified from SALS and FALS proband data can be assessed for their frequency, and whether any affect known ALS genes. Discordant CNVs may also be identified among

ALS-discordant MZ twin/triplet sets. Possible pathogenic CNVs can be filtered to discard any that are present within control datasets, including in-house and publicly available control cohorts before validating remaining CNVs. Any CNVs identified with compelling pathogenic potential can also be screened in extended patient cohorts.

8.6 Concluding remarks

In the 25 years since *SOD1* was discovered as the first known ALS gene, many advances have been made in our understanding of the genetic aetiology underlying ALS, with mutations in over 20 genes reported to cause disease. The strategies to identify ALS genes have also undergone many changes during this time. Earlier studies focussed on the analysis of large families using traditional genetic techniques such as genetic linkage analysis. Today, most of the ALS families with an unknown causal mutation are small and not amenable to traditional gentic linkage analysis, and are therefore analysed using next-generation sequencing strategies. Though the genetic discoveries in ALS to-date have provided significant advances in our understanding of disease pathogenesis, much is yet to be understood of the pathogenic mechanisms underlying ALS. Ongoing genetic discoveries offer a means of unravelling these mechanisms.

With the cause of disease yet to be identified in almost 40% of Australian FALS and over 90% of Australian SALS patients, many more mutations and genetic risk factors are yet to be discovered. The analyses presented in this thesis have made substantial progress towards identifying unknown genetic factors in ALS. The current genetic landscape of Australian familial ALS has been established, and the search for the ALS causal mutation in each of five Australian ALS families has been substantially narrowed, with just a handful of strong candidate mutations present in each. These results highlight the genetic heterogeneity of ALS, and the difficulties faced by ALS gene discovery research efforts. By shedding light on the complex nature of ALS genetics, the data generated here will provide vital guidance for the refinement of novel ALS gene discovery strategies moving forward.

The genetic analyses and biological insights into ALS presented in this thesis were made possible by the development of numerous bioinformatic scripting strategies. These strategies will continue to be utilised in the search for novel genetic causes of ALS using NGS data from established and expanding familial and sporadic patient cohorts.

Such NGS datasets are rich resources that will facilitate the identification of novel genetic causes of disease, and in turn allow the expansion of the genetic spectrum of ALS.

Novel ALS gene discoveries will continue to be vital resources to enhance our understanding of the mechanisms underlying disease pathogenesis. These novel gene mutations will form the basis of *in vitro* and *in vivo* models of disease, which will be required to not only decipher the intricate processes leading to disease onset and progression, but to also develop and test novel therapies. Novel gene discoveries will also have direct utility in the clinic by enabling diagnostic testing, carrier-testing and preimplantation genetic diagnosis. With no effective treatment or cure available for ALS, these genetic discoveries will give hope to patients and families affected by disease, that there is a future without ALS.

A Appendix

A.1 Ethics Approval

Appendix A.1 (Ethics Approval) of this thesis has been removed as it may contain sensitive/confidential content

A.2 Bioinformatics scripts

This section presents the complex bioinformatic scripts that were developed and/or used for genetics analysis of next-generation sequencing data as part of this thesis.

A.2.1 ANNOVAR annotation of the 850-sample WGS VCF

This code was used to subset annotate the 850-sample WGS VCF using with the information described in Table 2.2 using ANNOVAR software.

```
#!/bin/sh -e
#
#  VCF850_ANNOVAR.sh

#script to annotate full 850 WGS VCF

# run code from directory containing annovar .pl scripts
cd /datastore/mcc549/annovar

# make a new VCF with only meta data and sample 1 information
cut -f 1-10 complete.vcf | grep -v -P '^#' > sample1.vcf

# make a text file with sample information for all samples in columns 11-854
cut -f 11-859 complete.vcf | grep -v -P '^#' > samples11_854.vcf

# run annovar sample 1 VCF
./table_annovar.pl sample1.vcf humandb/ -buildver hg19 -out sample1_myanno
    -remove -protocol refGene,cytoBand,exac03,gnomad_exome,gnomad_genome,
    avsnp147,dbnsfp33a,dbnsfp31a_interpro,esp6500siv2_ea,esp6500siv2_all,
    ALL.sites.2015_08,EUR.sites.2015_08,clinvar_20170130 -operation
    g,r,f,f,f,f,f,f,f,f,f,f -nastring . -vcfinput

# paste everything back together
paste sample1_myanno.hg19_multianno.vcf samples11_854.vcf >
    ALLsamples_myanno.hg19_multianno.vcf

# make the header
grep '^#' complete.vcf > header.vcf

# add the header back in
```

```
26 cat header.vcf ALLsamples_myanno.hg19_multianno.vcf >
       ALLsamplesHEADED_myanno.hg19_multianno.vcf
27
28 # remove incomplete header
29 sed -e '2d' foo ALLsamplesHEADED_myanno.hg19_multianno.vcf >
       ALLsamplesFINAL_myanno.hg19_multianno.vcf
```

A.2.2 Family subsetting and removal of wild-type and uncalled variants

This code was used to subset an ALS family, FALSmq28, from the complete 850-sample WGS VCF, and subsequently remove variants which were uniformly not called or homozygous wild-type in all three family members.

```
#!/bin/sh
#
#  VCF850_family_subsetting.sh

# Take columns 1-9 for all lines after and including the line starting with a # symbol of a VCF, and write them to a new file
cut -f 1-9,836-838 EXAMPLE.vcf | grep -v -P '^#' > EXAMPLE_subset.vcf

# This script removes variants from 3-sample VCF that are either not called or homozygous widtype in all 3 individuals

# remove all variants with no genotype called in all 3 individuals (present in columns 10-12)
awk ' ! ($10 ~ /^\.\/\.:/ && $11 ~ /^\.\/\.:/ && $12 ~ /^\.\/\.:/) {print $0}' EXAMPLE.vcf > EXAMPLE_called.vcf

# remove all variants with a homozygous WT genotype in all 3 individuals (present in columns 10-12)
awk ' ! ($10 ~ /^0\/0:/ && $11 ~ /^0\/0:/ && $12 ~ /^0\/0:/) {print $0}' EXAMPLE_called.vcf > EXAMPLE_called_noWThom.vcf
```

A.2.3 Extracting allele count data from control database VCFs

These scripts were used to extract allele count data from very large control databases and append this data to patient VCFs for downstream comparisons. The R version was used for ExAC, gnomAD and Diamantina, and the SNPSift version was used for MGRB.

A.2.3.1 R version

```
# control_database_allele_count_data_Rversion.R

# This code is for extracting allele count data from a very large control database VCF and appending this data to patient VCFs for downstream comparisons with ALS patients
# As an example, the ExAC control database VCF is used here

# load required R libraries
library(VariantAnnotation)
library(BiocInstaller)

# see what fields are present in this VCF
scanVcfHeader("/Volumes/Emilly\ 1TB/ExAC.r0.3.1.sites.vep.vcf")

# define the paramaters on which we want to filter the vcf file
AC.adj.param <- ScanVcfParam(info="AC_Adj") # corrected alternate allele count
AN.adj.param <- ScanVcfParam(info="AN_Adj") # corrected total allele count


# load total allele counts (AN) and alt allele counts (AC) for all variants present in the control db VCF
raw.exac.AC.adj. <- readVcf("/Volumes/Emilly\ 1TB/ExAC.r0.3.1.sites.vep.vcf", "hg19", param=AC.adj.param) # s3 class object
raw.exac.AN.adj. <- readVcf("/Volumes/Emilly\ 1TB/ExAC.r0.3.1.sites.vep.vcf", "hg19", param=AN.adj.param) # s3 class object

```

```
# extract the the INFO column (AC and AN) data and genomic ranges information for each variant and combine
row.ranges.AC.adj <- rowRanges(raw.exac.AC.adj.)
mcols(row.ranges.AC.adj) <- info(raw.exac.AC.adj.)
row.ranges.AN.adj <- rowRanges(raw.exac.AN.adj.)
mcols(row.ranges.AN.adj) <- info(raw.exac.AN.adj.)

# make these data frames
df.exac.AC.adj <- as.data.frame(row.ranges.AC.adj)
df.exac.AN.adj <- as.data.frame(row.ranges.AN.adj)

# add chr.position column as an identifying column
df.exac.AC.adj$chr.position <- paste(df.exac.AC.adj$seqnames, df.exac.AC.adj$start, sep = ":")
df.exac.AN.adj$chr.position <- paste(df.exac.AN.adj$seqnames, df.exac.AN.adj$start, sep = ":")

# remove any unnecessary rows
row.names(df.exac.AC.adj) <- NULL
df.exac.AC.adj$seqnames <- NULL
df.exac.AC.adj$start <- NULL
df.exac.AC.adj$end <- NULL
df.exac.AC.adj$width <- NULL
df.exac.AC.adj$strand <- NULL

row.names(df.exac.AN.adj) <- NULL
df.exac.AN.adj$seqnames <- NULL
df.exac.AN.adj$start <- NULL
df.exac.AN.adj$end <- NULL
df.exac.AN.adj$width <- NULL
df.exac.AN.adj$strand <- NULL

# bring in annotated file of patient samples
patients <- read.delim("/Volumes/Personal//Bioinformatics/Files to work with/Brisbane_MND.Kelly.hg19_multianno.xls")

# add a chr:position column to the patients dataframe
patients$chr.position <- paste(patients$Chr, patients$Start, sep = ":")

```

```
# get rid of the "chr" so merging can occur
patients$chr.position <- sub("chr", "", patients$chr.position , perl=T)

# merge control db allele count data frames on to the end of the patients
    data frame matching on chr:position
patients <- merge( x=patients, y=df.exac.AN.adj, by.x="chr.position",
    by.y="chr.position", all.x = TRUE )
patients <- merge( x=patients, y=df.exac.AC.adj, by.x="chr.position",
    by.y="chr.position", all.x = TRUE )

# reorder variants back to ascending, by Chr and Start
patients <- patients[ order(patients$Chr, patients$Start), ]

# make allele counts numeric (default to character class which cannot be
    used for arithmetic or statistics downstream)
patients$AN_Adj <- as.numeric(patients$AN_Adj)
patients$AC_Adj <- sub("c\\S+", "x", patients$AC_Adj, perl=T)
patients$AC_Adj <- as.numeric(patients$AC_Adj)

# make a new column with control db ref allele counts calculated from AN -
    AC
patients$ref.allele.exac <- patients$AN_Adj-patients$AC_Adj

# rename exac allele count columns with more intuitive names
names(patients)[names(patients)=="AN_Adj"] <- "total.allele.exac"
names(patients)[names(patients)=="AC_Adj"] <- "alt.allele.exac"

# now save as an R object so we can use it elsewhere
save(patients, file = "FALS_WES_EXAC.RObject")
```

A.2.3.2 SNPSift version

```
# control_database_allele_count_data_SNPSiftversion.R

# This code is for extracting allele count data from a very large control database VCF and appending this data to patient VCFs for downstream comparisons with ALS patients
# As an example, the MGRB control database VCF is used here

# load required R libraries
library(VariantAnnotation)
library(BiocInstaller)
library(splitstackshape)

# see what fields are present in this VCF
scanVcfHeader("/Volumes/Emilly\ 1TB/MGRB.vcf")

# Use SNPSift in UNIX to extract meta and allele count information - extractFields_VCF.sh

# bring in MGRB allele count data
MGRB <- read.delim("MGRB_allele_data.txt")

# add a chr:position column to the patients dataframe
MGRB$chr.position <- paste(MGRB$Chr, MGRB$Start, sep = ":")

# split INFO columns
MGRB_df <- cSplit(MGRB, c("ALT", "AC", "AF"), sep=",", type.convert=TRUE)

# create total allele count column
MGRB_df$AN <- 2*(MGRB_df$NS)

# create REF allele count column
MGRB_df$AR <- MGRB_df$AN-MGRB_df$AC_1

# subset allele count data for ease of merging downstream
MGRB_df_small <- cbind(MGRB_df$exact.position, MGRB_df$AN, MGRB_df$AR, MGRB_df$AC_1)

```

```
# rename columns
colnames(MGRB_df_small) <- c("exact.position", "AN_welderly",
    "AR_welderly", "AC_welderly")

# bring in annotated file of patient samples
patients <- read.delim("/Volumes/Personal//Bioinformatics/Files to work
    with/Brisbane_MND.Kelly.hg19_multianno.xls")

# add a chr:position column to the patients dataframe
patients$chr.position <- paste(patients$Chr, patients$Start, sep = ":")

# get rid of the "chr" so merging can occur
patients$chr.position <- sub("chr", "", patients$chr.position , perl=T)

# merge MGRB allele count data frames on to the end of the patients data
    frame matching on chr:position
patients <- merge( x=patients, y=MGRB_df_small, by.x="chr.position",
    by.y="chr.position", all.x = TRUE )

# reorder variants back to ascending, by Chr and Start
patients <- patients[ order(patients$Chr, patients$Start), ]

# make allele counts numeric (default to character class which cannot be
    used for arithmetic or statistics downstream)
patients$AN_welderly <- as.numeric(AN_welderly)
patients$AN_welderly <- as.numeric(AR_welderly)
patients$AN__welderly <- as.numeric(AC_welderly)

# now save as an R object so we can use it elsewhere
save(patients, file = "FALS_WES_MGRB.RObject")
```

A.2.4 Candidate gene searching and association analysis in FALS WES data

This R markdown code was used to identify variants in candidate ALS genes among WES data from different cohorts of FALS patients/family members, and perform association testing on known SNPs using allele count data from patients and ExAC controls (or MGRB or Diamantina controls).

```
---
title: "FALS_WES_candidate_gene_analysis.Rmd"
output: html_document
---

# This code searches all MQ exome data (and subsets thereof), appended with ExAC allele counts, for a given set of genes, and then ouputs all variants identified in MQ exome data for this gene across all samples, tallies genotypes and allele frequencies and perfroms a Fishers exact test comparing allele frequencies between ExAC controls and MQ patients for each variant.

Preface i. Load required packages and setting directories
```{r install.libraries, cache=FALSE}
library(BiocInstaller)
library(WriteXLS)
library(dplyr)
library(data.table)
```

```{r set.directories, cache=FALSE}
data="/Volumes/Personal/Bioinformatics/Candidate\ gene\ hunting\ in\ R/Candidate_gene_hunting/Raw\ data"
directory="/Volumes/Personal/Bioinformatics/Candidate\ gene\ hunting\ in\ R/Candidate_gene_hunting/QC\ and\ analysis"

#source="/Volumes/Personal/Bioinformatics/Candidate\ gene\ hunting\ in\ R/Candidate_gene_hunting/functions"
```

Preface ii. Importing files and getting them ready to work with
```{r import.files, cache=TRUE}
```
```

```
25
26 setwd(directory)
27
28 # import and view annotated file of all samples with exac allele counts
29 load("/Volumes/Personal/Bioinformatics/Candidate\ gene\ hunting\ in\
       R/Candidate_gene_hunting/QC\ and\
       analysis/all.samples.all.annotated.variants.RObject")
30 View(all.samples.all.annotated.variants)
31 ```
32
33 # Section 1. Search for known ALS gene variants
34 ```{r known.genes, cache=FALSE}
35 # look for variants in the known ALS genes and pull out all associtaed info
       (ie. the entire row for eacch variant of the candidate gene)
36
37 ## first define known genes
38 ### ALSOD ALS genes and CCNF
39 x <- c("SOD1", "ALS2", "ALS3", "SETX", "SPG11", "FUS", "ALS7", "VAPB",
       "ANG", "TARDBP", "FIG4", "OPTN", "ATXN2", "VCP", "UBQLN2", "SIGMAR1",
       "CHMP2B", "PFN1", "ERBB4", "HNRNPA1", "MATR3", "CHCHD10", "C9orf72",
       "UNC13A", "DAO", "DCTN1", "NEFH", "PRPH", "SQSTM1", "TAF15", "SPAST",
       "ELP3", "LMNB1", "CCNF")
40
41 #### run all known ALS genes through all samples and output a file
       containing all ALS gene variants
42 ALS.gene.variants <- subset(all.samples.all.annotated.variants,
       Gene.refGene %in% x)
43
44 ## first define known genes
45 ### ALSOD other genes
```

```
y <- c("APEX1", "APOE", "AR", "CCS", "CNTF", "CYP2D6", "ALAD", "DYNC1H1",
    "CHGB", "GLE1", "TBK1", "ITPR2", "GRN", "LIPC", "NT5C1A", "ZFP64",
    "DISC1", "SLC39A11", "ZNF746", "FGGY", "DPP6", "LUM", "RNF19A", "SOX5",
    "OMA1", "GRB14", "PON2", "PON3", "SLC1A2", "SMN1", "SMN2", "SNCG",
    "SUSD1", "B4GALT6", "OGG1", "AGT", "C1orf27", "VPS54", "FEZF2", "DOC2B",
    "CNTN6", "PSEN1", "PVR", "SOD2", "SPG7", "VDR", "VEGFA", "RBMS1",
    "CSNK1G3", "BCL11B", "NETO1", "CDH22", "DIAPH3", "GARS", "HEXA", "HFE",
    "KIFAP3", "LIF", "LOX", "MAOB", "MAPT", "MT-ND2", "NAIP", "CRYM",
    "SYT9", "CRIM1", "SCN7A", "EFEMP1", "KDR", "CDH13", "PON1", "CNTN4",
    "SELL", "EWSR1", "PARK7", "HNRNPA2B1", "NIPA1", "SEMA6A", "ZNF512B",
    "RNASE2", "PLEKHG5", "BCL6", "RAMP3", "SS18L1", "PCP4", "CST3", "EPHA4",
    "ARHGEF28", "TRPM7", "SARM1", "CX3CR1", "TUBA4A", "SYNE")

#### run all other ALSOD genes through all samples and output a file
    containing all other ALSOD gene variants
ALS.associated.gene.variants <- subset(all.samples.all.annotated.variants,
    Gene.refGene %in% y)
```

```
# Section 2. Patient cohort subsetting
```{r subset.exomes, cache=TRUE}
# subet combined exomes to patient cohorts

## define coloumn/sample names of patients to be excluded (-c(...)) or
    included (c(...))
### FALS with an unidentified ALS mutation
unknown.mut <- subset(all.samples.all.annotated.variants, , -c(X10.000094,
    X10.000094.2015, X10.020768, X10.040672, X10.971251, X10.971251.2015,
    X119.020807, X119.050471, X119.050612, X119.960341, X13.080638,
    X13.090339, X13.A188, X13.A203, X13.A217, X14.950280, X147.020183,
    X166.000485, X171.030781, X187.050743, X187.100285, X187.960560,
    X194.060051, X270.090391, X285.100287, X291.100225, X304.100786,
    X32.940107, X32.940634, X32.A269, X45.040247, X45.040542, X5.100248,
    X5.970585, X5.A124, X51.A348, X6.010076, X6.A135, X67.040088,
    X67.960247, X73.100120, X73.100121, X73.940147, X77.040435, X77.940361,
    X77.950083, X82.940727, X86.030814, X86.950057, X86.950164, X86.950165,
    X92.950426, X92.950427, X92.950430, MQ130084, mq1.MQ140002))

### All FALS patients and obligate carriers
```
```

```
aff.ob <- subset(all.samples.all.annotated.variants, , -c(X10.020768,
    X10.040672, X119.050612, X147.020183, X166.000485, X187.100285,
    X194.060051, X45.040247, X45.040542, X51.A348, X73.940147, X92.950427,
    mq1.MQ140002))

### All FALS with a pathogenic expansion in C9orf72 (identified by repeat
    primed PCR after WES)
C9.pos <- subset(all.samples.all.annotated.variants, , c(1:68, X119.020807,
    X119.050471, X119.960341, X13.080638, X13.090339, X13.A188, X13.A203,
    X13.A217, X187.050743, X187.960560, X291.100225, X32.940107, X32.940634,
    X32.A269, X6.010076, X6.A135, X67.040088, X67.960247, X73.100120,
    X73.100121, X77.040435, X77.940361, X77.950083, X86.030814, X86.950057,
    X86.950164, X86.950165, X92.950426, X92.950427, X92.950430, 206:208))
```

```

# Section 3. Search for candidate ALS gene variants
```{r candidate.genes, cache=FALSE}
# look for variants in the candidate genes and pull out all associtaed info
    (ie. the entire row for eacch variant of the candidate gene)
# this will be performed in each patient cohort subset of exomes
    (unknown.mut, aff.ob, C9.pos)

## first define cadidate genes
### for example (actual code contains each candidate gene search set and
    the date analysis was performed)
z <- c("PURA", "NEK1", "C21orf2", "MOBP", "SCFD1", "SPTBN4")

#### run all other ALSOD genes through all samples and output a file
    containing all other ALSOD gene variants
all.candidate.gene.variants <- subset(all.samples.all.annotated.variants,
    Gene.refGene %in% z)
unknown.mut.candidate.gene.variants <- subset(unknown.mut, Gene.refGene
    %in% z)
aff.ob.candidate.gene.variants <- subset(aff.ob, Gene.refGene %in% z)
C9.pos.candidate.gene.variants <- subset(C9.pos, Gene.refGene %in% z)
```

```

# Section 4. Genotype counting
```{r genotype.counting, cache=FALSE}
```

```
85 # add columns containing counts of how many patients have the given
       genotype for each variant
86 ## All FALS WES data
87 all.candidate.gene.variants$No.hom_patients <-
       apply(all.candidate.gene.variants[,69:205], 1, function(u)
       length(which(grepl("1/1",u))==TRUE) )
88 all.candidate.gene.variants$No.het_patients <-
       apply(all.candidate.gene.variants[,69:205], 1, function(u)
       length(which(grepl("0/1|1/0",u))==TRUE) )
89 all.candidate.gene.variants$No.WT_patients <-
       apply(all.candidate.gene.variants[,69:205], 1, function(u)
       length(which(grepl("0/0",u))==TRUE) )
90
91 ## unknown.mut FALS patient cohort subset
92 unknown.mut.candidate.gene.variants$No.hom_patients <-
       apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
       length(which(grepl("1/1",u))==TRUE) )
93 unknown.mut.candidate.gene.variants$No.het_patients <-
       apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
       length(which(grepl("0/1|1/0",u))==TRUE) )
94 unknown.mut.candidate.gene.variants$No.WT_patients <-
       apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
       length(which(grepl("0/0",u))==TRUE) )
95
96 ## aff.ob FALS patient cohort subset
97 aff.ob.candidate.gene.variants$No.hom_patients <-
       apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
       length(which(grepl("1/1",u))==TRUE) )
98 aff.ob.candidate.gene.variants$No.het_patients <-
       apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
       length(which(grepl("0/1|1/0",u))==TRUE) )
99 aff.ob.candidate.gene.variants$No.WT_patients <-
       apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
       length(which(grepl("0/0",u))==TRUE) )
100
101 ## C9.pos FALS patient cohort subset
102 C9.pos.candidate.gene.variants$No.hom_patients <-
       apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
       length(which(grepl("1/1",u))==TRUE) )
```

```
C9.pos.candidate.gene.variants$No.het_patients <-
    apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) )
C9.pos.candidate.gene.variants$No.WT_patients <-
    apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )
```
```

# Section 5. Allele counting
```{r allele.counting, cache=FALSE}
# Calculate patient allele counts for each variant based on the genotype
    columns, and add allele count columns
## All FALS WES data
all.candidate.gene.variants$Patient_alt_allele_count <- (
    2*(apply(all.candidate.gene.variants[,69:205], 1, function(u)
    length(which(grepl("1/1",u))==TRUE) )) +
    apply(all.candidate.gene.variants[,69:205], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
all.candidate.gene.variants$Patient_ref_allele_count <- (
    2*(apply(all.candidate.gene.variants[,69:205], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )) +
    apply(all.candidate.gene.variants[,69:205], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
all.candidate.gene.variants$Patient_total_allele_count <-
    all.candidate.gene.variants$Patient_alt_allele_count +
    all.candidate.gene.variants$Patient_ref_allele_count

## unknown.mut FALS patient cohort subset
unknown.mut.candidate.gene.variants$Patient_alt_allele_count <- (
    2*(apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
    length(which(grepl("1/1",u))==TRUE) )) +
    apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
unknown.mut.candidate.gene.variants$Patient_ref_allele_count <- (
    2*(apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )) +
    apply(unknown.mut.candidate.gene.variants[,69:149], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
```
```

```
unknown.mut.candidate.gene.variants$Total_Patient_allele_count <-
    unknown.mut.candidate.gene.variants$Patient_ref_allele_count +
    unknown.mut.candidate.gene.variants$Patient_alt_allele_count

## aff.ob FALS patient cohort subset
aff.ob.candidate.gene.variants$Patient_alt_allele_count <- (
    2*(apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
    length(which(grepl("1/1",u))==TRUE) )) +
    apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
aff.ob.candidate.gene.variants$Patient_ref_allele_count <- (
    2*(apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )) +
    apply(aff.ob.candidate.gene.variants[,69:192], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
aff.ob.candidate.gene.variants$Total_Patient_allele_count <-
    aff.ob.candidate.gene.variants$Patient_ref_allele_count +
    aff.ob.candidate.gene.variants$Patient_alt_allele_count

## C9.pos FALS patient cohort subset
C9.pos.candidate.gene.variants$Patient_alt_allele_count <- (
    2*(apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("1/1",u))==TRUE) )) +
    apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
C9.pos.candidate.gene.variants$Patient_ref_allele_count <- (
    2*(apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )) +
    apply(C9.pos.candidate.gene.variants[,69:98], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ) )
C9.pos.candidate.gene.variants$Total_Patient_allele_count <-
    C9.pos.candidate.gene.variants$Patient_ref_allele_count +
    C9.pos.candidate.gene.variants$Patient_alt_allele_count
```` ``` ````

```
# Section 6. Association testing
```` ```{r fishers.exact, cache=FALSE} ````
Perform Fisher's exact testing comparing patient and ExAC allele counts,
    and add a new column containing the resultant p-value
```
```

```
## All FALS WES data
res <- NULL
for (i in 1:nrow(all.candidate.gene.variants)){
  table <- matrix(c(all.candidate.gene.variants[i,213],
      all.candidate.gene.variants[i,212],
      all.candidate.gene.variants[i,208],
      all.candidate.gene.variants[i,207]), ncol = 2, byrow = TRUE)
  # if any NA occurs in your table save an error in p else run the fisher
      test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
  res <- c(res,p)
}
all.candidate.gene.variants$fishers <- res


## unknown.mut FALS patient cohort subset
res <- NULL
for (i in 1:nrow(unknown.mut.candidate.gene.variants)){
  table <- matrix(c(unknown.mut.candidate.gene.variants[i,157],
      unknown.mut.candidate.gene.variants[i,156],
      unknown.mut.candidate.gene.variants[i,152],
      unknown.mut.candidate.gene.variants[i,151]), ncol = 2, byrow = TRUE)
  # if any NA occurs in your table save an error in p else run the fisher
      test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
  res <- c(res,p)
}
unknown.mut.candidate.gene.variants$fishers <- res


## aff.ob FALS patient cohort subset
res <- NULL
for (i in 1:nrow(aff.ob.candidate.gene.variants)){
  table <- matrix(c(aff.ob.candidate.gene.variants[i,200],
      aff.ob.candidate.gene.variants[i,199],
      aff.ob.candidate.gene.variants[i,195],
      aff.ob.candidate.gene.variants[i,194]), ncol = 2, byrow = TRUE)
```

```
  # if any NA occurs in your table save an error in p else run the fisher test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
  res <- c(res,p)
}
aff.ob.candidate.gene.variants$fishers <- res


## C9.pos FALS patient cohort subset
res <- NULL
for (i in 1:nrow(C9.pos.candidate.gene.variants)){
  table <- matrix(c(C9.pos.candidate.gene.variants[i,106],
      C9.pos.candidate.gene.variants[i,105],
      C9.pos.candidate.gene.variants[i,101],
      C9.pos.candidate.gene.variants[i,100]), ncol = 2, byrow = TRUE)
  # if any NA occurs in your table save an error in p else run the fisher test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
  res <- c(res,p)
}
C9.pos.candidate.gene.variants$fishers <- res
```

A.2.5 Association analysis for all possible family combinations in FALS WES data

This script was used to perform association testing for a single SNP using Fisher's exact testing for each possible combination including a single member of each family with an unknown ALS causal mutation.

```
# FALS_assoc_SNP_family_loop.R

# This script takes a candidate gene variant found to be associated with
    disease in all 81 FALS individuals with an unknown ALS mutation
    (identified using FALS_WES_candidate_gene_analysis.Rmd), and performs
    fisher's exact testing for all possible combinations of FALS individuals
    with an unknown ALS mutation where only a single member of each FALS
    family is included

# subset the the associated variant from the candidate gene dataframe
## for example, 12:57969016
candidate.variant <- subset(unknown.mut.candidate.gene.variants,
    chr.position == "12:57969016")

# delete patient genotype and allele counts calculated from all unknown
    exomes
# (which includes multiple individuals from some families)
candidate.variant$No.hom_patients <- NULL
candidate.variant$No.het_patients <- NULL
candidate.variant$No.WT_patients <- NULL
candidate.variant$Patient_alt_allele_count <- NULL
candidate.variant$Patient_ref_allele_count <- NULL
candidate.variant$Total_Patient_allele_count <- NULL
candidate.variant$fishers <- NULL

# append the family identifiers for each sample to the END of the dataframe
df <- as.data.frame(c(candidate.variant, "FALS100", "FALS101", "FALS115",
    "FALS116", "FALS117", "FALS122", "FALS136", "FALS143", "FALS145",
    "FALS147", "FALS147", "FALS15", "FALS15", "FALS151", "FALS153",
    "FALS154", "FALS158", "FALS159", "FALS162", "FALS172","FALS175",
    "FALS176", "FALS180", "FALS181", "FALS184", "FALS185", "FALS196",
    "FALS199", "FALS205", "FALS206", "FALS206", "FALS206", "FALS206",
    "FALS215","FALS239", "FALS245", "FALS251", "FALS280", "FALS283",
```

```
"FALS29", "FALS292", "FALS292", "FALS294", "FALS296", "FALS300", "FALS302",
    "FALS303", "FALS305", "FALS307", "FALS314", "FALS318", "FALS321",
    "FALS326", "FALS328", "FALS40", "FALS42", "FALS45", "FALS63", "FALS71",
    "FALS79", "FALS8", "FALS8", "FALS8", "FALS93", "mq21", "mq20", "mq2",
    "mq4", "mq15", "mq12", "mq13", "mq22", "mq14", "mq17", "mq18", "mq19",
    "mq1", "mq1", "mq1", "mq2", "mq20"))

# reorder the df so that ExAC allele counts are ahead of our sample and
    family information
df <- df[ ,c(1:68,150:152,69:149,153:233)]

# ensure df is in character format
df[] <- lapply(df, as.character)

# find the postions of each family identifier data and assign to
    appropriate variables
# for those families with muliple individuals, there will be multiple
    positions
FALS100.var <- which(df =="FALS100")
FALS101.var <- which(df =="FALS101")
FALS115.var <- which(df =="FALS115")
FALS116.var <- which(df =="FALS116")
FALS117.var <- which(df =="FALS117")
FALS122.var <- which(df =="FALS122")
FALS136.var <- which(df =="FALS136")
FALS143.var <- which(df =="FALS143")
FALS145.var <- which(df =="FALS145")
FALS147.var <- which(df =="FALS147")
FALS15.var <- which(df =="FALS15")
FALS151.var <- which(df =="FALS151")
FALS153.var <- which(df =="FALS153")
FALS154.var <- which(df =="FALS154")
FALS158.var <- which(df =="FALS158")
FALS159.var <- which(df =="FALS159")
FALS162.var <- which(df =="FALS162")
FALS172.var <- which(df =="FALS172")
FALS175.var <- which(df =="FALS175")
FALS176.var <- which(df =="FALS176")
FALS180.var <- which(df =="FALS180")
```

```
FALS181.var <- which(df =="FALS181")
FALS184.var <- which(df =="FALS184")
FALS185.var <- which(df =="FALS185")
FALS196.var <- which(df =="FALS196")
FALS199.var <- which(df =="FALS199")
FALS205.var <- which(df =="FALS205")
FALS206.var <- which(df =="FALS206")
FALS215.var <- which(df =="FALS215")
FALS239.var <- which(df =="FALS239")
FALS245.var <- which(df =="FALS245")
FALS251.var <- which(df =="FALS251")
FALS280.var <- which(df =="FALS280")
FALS283.var <- which(df =="FALS283")
FALS29.var <- which(df =="FALS29")
FALS292.var <- which(df =="FALS292")
FALS294.var <- which(df =="FALS294")
FALS296.var <- which(df =="FALS296")
FALS300.var <- which(df =="FALS300")
FALS302.var <- which(df =="FALS302")
FALS303.var <- which(df =="FALS303")
FALS305.var <- which(df =="FALS305")
FALS307.var <- which(df =="FALS307")
FALS314.var <- which(df =="FALS314")
FALS318.var <- which(df =="FALS318")
FALS321.var <- which(df =="FALS321")
FALS326.var <- which(df =="FALS326")
FALS328.var <- which(df =="FALS328")
FALS40.var <- which(df =="FALS40")
FALS42.var <- which(df =="FALS42")
FALS45.var <- which(df =="FALS45")
FALS63.var <- which(df =="FALS63")
FALS71.var <- which(df =="FALS71")
FALS79.var <- which(df =="FALS79")
FALS8.var <- which(df =="FALS8")
FALS93.var <- which(df =="FALS93")
mq21.var <- which(df =="mq21")
mq20.var <- which(df =="mq20")
mq2.var <- which(df =="mq2")
mq4.var <- which(df =="mq4")
```

```
mq15.var <- which(df =="mq15")
mq12.var <- which(df =="mq12")
mq13.var <- which(df =="mq13")
mq22.var <- which(df =="mq22")
mq14.var <- which(df =="mq14")
mq17.var <- which(df =="mq17")
mq18.var <- which(df =="mq18")
mq19.var <- which(df =="mq19")
mq1.var <- which(df =="mq1")

# get all possible position combinations including one of each family
    identifier
family.combinations <- expand.grid(FALS100.var, FALS101.var, FALS115.var,
    FALS116.var, FALS117.var, FALS122.var, FALS136.var, FALS143.var,
    FALS145.var, FALS147.var, FALS15.var, FALS151.var, FALS153.var,
    FALS154.var, FALS158.var, FALS159.var, FALS162.var, FALS172.var,
    FALS175.var, FALS176.var, FALS180.var, FALS181.var, FALS184.var,
    FALS185.var, FALS196.var, FALS199.var, FALS205.var, FALS206.var,
    FALS215.var, FALS239.var, FALS245.var, FALS251.var, FALS280.var,
    FALS283.var, FALS29.var, FALS292.var, FALS294.var, FALS296.var,
    FALS300.var, FALS302.var, FALS303.var, FALS305.var, FALS307.var,
    FALS314.var, FALS318.var, FALS321.var, FALS326.var, FALS328.var,
    FALS40.var, FALS42.var, FALS45.var, FALS63.var, FALS71.var, FALS79.var,
    FALS8.var, FALS93.var, mq21.var, mq20.var, mq2.var, mq4.var, mq15.var,
    mq12.var, mq13.var, mq22.var, mq14.var, mq17.var, mq18.var, mq19.var,
    mq1.var)

# provide correct names
names(family.combinations) <- c("FALS100", "FALS101", "FALS115", "FALS116",
    "FALS117", "FALS122", "FALS136", "FALS143", "FALS145", "FALS147",
    "FALS15", "FALS151", "FALS153", "FALS154", "FALS158", "FALS159",
    "FALS162", "FALS172","FALS175", "FALS176", "FALS180", "FALS181",
    "FALS184", "FALS185", "FALS196", "FALS199", "FALS205", "FALS206",
    "FALS215", "FALS239", "FALS245", "FALS251", "FALS280", "FALS283",
    "FALS29", "FALS292", "FALS294", "FALS296", "FALS300", "FALS302",
    "FALS303", "FALS305", "FALS307", "FALS314", "FALS318", "FALS321",
    "FALS326", "FALS328", "FALS40", "FALS42", "FALS45", "FALS63", "FALS71",
    "FALS79", "FALS8", "FALS93", "mq21", "mq20", "mq2", "mq4", "mq15",
    "mq12", "mq13", "mq22", "mq14", "mq17", "mq18", "mq19", "mq1")
```

```
106
107 # create a results data frame
108 df.combinations <- as.data.frame(matrix(NA,ncol = 215, nrow =
        nrow(family.combinations)))
109
110 # name the variables (columns) in the results dataframe
111 names(df.combinations) <- c("chr.position", "Chr", "Start", "End", "Ref",
        "Alt", "Func.refGene", "Gene.refGene", "GeneDetail.refGene",
        "ExonicFunc.refGene", "AAChange.refGene", "gerp..elem",
        "phastConsElements46way", "genomicSuperDups", "ExAC_ALL", "ExAC_AFR",
        "ExAC_AMR", "ExAC_EAS", "ExAC_FIN", "ExAC_NFE", "ExAC_OTH", "ExAC_SAS",
        "esp6500si_all", "esp6500si_aa", "esp6500si_ea", "X1000g2014oct_all",
        "X1000g2014oct_eur", "X1000g2014oct_amr", "X1000g2014oct_asn",
        "X1000g2014oct_afr", "snp129", "avsnp142", "clinvar_20150330", "avsift",
        "SIFT_score", "SIFT_pred", "Polyphen2_HDIV_score",
        "Polyphen2_HDIV_pred", "Polyphen2_HVAR_score", "Polyphen2_HVAR_pred",
        "LRT_score", "LRT_pred", "MutationTaster_score", "MutationTaster_pred",
        "MutationAssessor_score", "MutationAssessor_pred", "FATHMM_score",
        "FATHMM_pred", "RadialSVM_score", "RadialSVM_pred", "LR_score",
        "LR_pred", "VEST3_score", "CADD_raw", "CADD_phred", "GERP..._RS",
        "phyloP46way_placental", "phyloP100way_vertebrate",
        "SiPhy_29way_logOdds", "CHROM", "POS", "ID", "REF", "ALT", "QUAL",
        "FILTER", "INFO", "FORMAT", "total.allele.exac", "alt.allele.exac",
        "ref.allele.exac",
112                         "FALS100.sample.pos", "FALS100.result",
113                         "FALS101.sample.pos", "FALS101.result",
114                         "FALS115.sample.pos", "FALS115.result",
115                         "FALS116.sample.pos", "FALS116.result",
116                         "FALS117.sample.pos", "FALS117.result",
117                         "FALS122.sample.pos", "FALS122.result",
118                         "FALS136.sample.pos", "FALS136.result",
119                         "FALS143.sample.pos", "FALS143.result",
120                         "FALS145.sample.pos", "FALS145.result",
121                         "FALS147.sample.pos", "FALS147.result",
122                         "FALS15.sample.pos", "FALS15.result",
123                         "FALS151.sample.pos", "FALS151.result",
124                         "FALS153.sample.pos", "FALS153.result",
125                         "FALS154.sample.pos", "FALS154.result",
126                         "FALS158.sample.pos", "FALS158.result",
```

```
                    "FALS159.sample.pos", "FALS159.result",
                    "FALS162.sample.pos", "FALS162.result",
                    "FALS172.sample.pos", "FALS172.result",
                    "FALS175.sample.pos", "FALS175.result",
                    "FALS176.sample.pos", "FALS176.result",
                    "FALS180.sample.pos", "FALS180.result",
                    "FALS181.sample.pos", "FALS181.result",
                    "FALS184.sample.pos", "FALS184.result",
                    "FALS185.sample.pos", "FALS185.result",
                    "FALS196.sample.pos", "FALS196.result",
                    "FALS199.sample.pos", "FALS199.result",
                    "FALS205.sample.pos", "FALS205.result",
                    "FALS206.sample.pos", "FALS206.result",
                    "FALS215.sample.pos", "FALS215.result",
                    "FALS239.sample.pos", "FALS239.result",
                    "FALS245.sample.pos", "FALS245.result",
                    "FALS251.sample.pos", "FALS251.result",
                    "FALS280.sample.pos", "FALS280.result",
                    "FALS283.sample.pos", "FALS283.result",
                    "FALS29.sample.pos", "FALS29.result",
                    "FALS292.sample.pos", "FALS292.result",
                    "FALS294.sample.pos", "FALS294.result",
                    "FALS296.sample.pos", "FALS296.result",
                    "FALS300.sample.pos", "FALS300.result",
                    "FALS302.sample.pos", "FALS302.result",
                    "FALS303.sample.pos", "FALS303.result",
                    "FALS305.sample.pos", "FALS305.result",
                    "FALS307.sample.pos", "FALS307.result",
                    "FALS314.sample.pos", "FALS314.result",
                    "FALS318.sample.pos", "FALS318.result",
                    "FALS321.sample.pos", "FALS321.result",
                    "FALS326.sample.pos", "FALS326.result",
                    "FALS328.sample.pos", "FALS328.result",
                    "FALS40.sample.pos", "FALS40.result",
                    "FALS42.sample.pos", "FALS42.result",
                    "FALS45.sample.pos", "FALS45.result",
                    "FALS63.sample.pos", "FALS63.result",
                    "FALS71.sample.pos", "FALS71.result",
                    "FALS79.sample.pos", "FALS79.result",
```

```
                        "FALS8.sample.pos", "FALS8.result",
                        "FALS93.sample.pos", "FALS93.result",
                        "mq21.sample.pos", "mq21.result",
                        "mq20.sample.pos", "mq20.result",
                        "mq2.sample.pos", "mq2.result",
                        "mq4.sample.pos", "mq4.result",
                        "mq15.sample.pos", "mq15.result",
                        "mq12.sample.pos", "mq12.result",
                        "mq13.sample.pos", "mq13.result",
                        "mq22.sample.pos", "mq22.result",
                        "mq14.sample.pos", "mq14.result",
                        "mq17.sample.pos", "mq17.result",
                        "mq18.sample.pos", "mq18.result",
                        "mq19.sample.pos", "mq19.result",
                        "mq1.sample.pos", "mq1.result",
                        "patient.WT","patient.het","patient.hom",
                        "patient.ref.count","patient.alt.count",
                        "fishers")

# copy in common data (common to each combination (row) ie chr position,
    exac allele counts etc)
df.combinations[,1:71] <- df[,1:71]

# setup variables based on combination data
for(i in 1:nrow(family.combinations)){
  df.combinations[i,c(72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96,
      98, 100, 102, 104, 106, 108, 110, 112, 114, 116, 118, 120, 122, 124,
      126, 128, 130, 132, 134, 136, 138, 140, 142, 144, 146, 148, 150, 152,
      154, 156, 158, 160, 162, 164, 166, 168, 170, 172, 174, 176, 178, 180,
      182, 184, 186, 188, 190, 192, 194, 196, 198, 200, 202, 204, 206, 208)]
      <- family.combinations[i,]
  # -81 to correct for the position of the results not the 'family type'
      data (number of family numbers appended at beginning)
  e.cycle.results <- as.numeric(family.combinations[i,] -81)
  df.combinations[i,c(73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97,
      99, 101, 103, 105, 107, 109, 111, 113, 115, 117, 119, 121, 123, 125,
      127, 129, 131, 133, 135, 137, 139, 141, 143, 145, 147, 149, 151, 153,
      155, 157, 159, 161, 163, 165, 167, 169, 171, 173, 175, 177, 179, 181,
      183, 185, 187, 189, 191, 193, 195, 197, 199, 201, 203, 205, 207, 209)]
```

```
194   <- df[e.cycle.results] }
195
196 # count patient genotypes
197 df.combinations$patient.WT <- apply(df.combinations[,c(73, 75, 77, 79, 81,
      83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109, 111, 113,
      115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 139, 141,
      143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163, 165, 167, 169,
      171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193, 195, 197,
      199, 201, 203, 205, 207, 209)], 1, function(u)
      length(which(grepl("0/0",u))==TRUE) )
198 df.combinations$patient.het <- apply(df.combinations[,c(73, 75, 77, 79, 81,
      83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109, 111, 113,
      115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 139, 141,
      143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163, 165, 167, 169,
      171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193, 195, 197,
      199, 201, 203, 205, 207, 209)], 1, function(u)
      length(which(grepl("0/1|1/0",u))==TRUE) )
199 df.combinations$patient.hom <- apply(df.combinations[,c(73, 75, 77, 79, 81,
      83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109, 111, 113,
      115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137, 139, 141,
      143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163, 165, 167, 169,
      171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193, 195, 197,
      199, 201, 203, 205, 207, 209)], 1, function(u)
      length(which(grepl("1/1",u))==TRUE) )
200
201 # count patient alleles
202 df.combinations$patient.ref.count <- ( 2*(apply(df.combinations[,c(73, 75,
      77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109,
      111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137,
      139, 141, 143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163, 165,
      167, 169, 171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193,
      195, 197, 199, 201, 203, 205, 207, 209)], 1, function(u)
    length(which(grepl("0/0",u))==TRUE) )) + apply(df.combinations[,c(73, 75,
        77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107,
        109, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135,
        137, 139, 141, 143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163,
        165, 167, 169, 171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191,
        193, 195, 197, 199, 201, 203, 205, 207, 209)], 1, function(u)
        length(which(grepl("0/1|1/0",u))==TRUE) ))
```

```
df.combinations$patient.alt.count <- ( 2*(apply(df.combinations[,c(73, 75,
    77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107, 109,
    111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135, 137,
    139, 141, 143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163, 165,
    167, 169, 171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 193,
    195, 197, 199, 201, 203, 205, 207, 209)], 1, function(u)
    length(which(grepl("1/1",u))==TRUE) )) + apply(df.combinations[,c(73,
    75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99, 101, 103, 105, 107,
    109, 111, 113, 115, 117, 119, 121, 123, 125, 127, 129, 131, 133, 135,
    137, 139, 141, 143, 145, 147, 149, 151, 153, 155, 157, 159, 161, 163,
    165, 167, 169, 171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191,
    193, 195, 197, 199, 201, 203, 205, 207, 209)], 1, function(u)
    length(which(grepl("0/1|1/0",u))==TRUE) ))

# perform fishers exact tests
res <- NULL
for (i in 1:nrow(df.combinations)){
  table <- matrix(as.numeric(c(df.combinations[i, 71], df.combinations[i,
      70], df.combinations[i, 213], df.combinations[i, 214])), ncol = 2,
      byrow = TRUE)
  # if any NA occurs in your table save an error in p else run the fisher
      test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
  res <- c(res,p)
}
df.combinations$fishers <- res

# add the results to the results dataframe set up eariler
df.combinations.results <-
    as.data.frame(cbind(1:nrow(df.combinations),df.combinations$fishers))
names(df.combinations.results) <- c("combo","fishers")
```

A.2.6 Candidate gene screening of the 850-sample WGS VCF

This script was used to parse the 850-sample WGS VCF for variants in any given gene, output these to a new file and add the appropriate column header information.

```
#!/bin/sh
#
#  850VCF_gene_search.sh

#  this code is for looking a specified gene in the 850VCF

# navigate to directory
cd /datastore/d/MND_Genomes/mcc549/candidate_genes

# define 850VCF
VCF850=/datastore/mcc549/annovar/myanno_ALS_Cohort.vqsr.vcf

# define gene name to be search for
GENE=CHCHD10

# define output file names
OUT1="$GENE"_"variants.vcf"
OUT2="$GENE"_"variants_headed.vcf"
OUT3="$GENE"_"variants_headed.txt"

# perform gene search
awk -v NAME="$GENE" '$8 ~ NAME { print $0 }' $VCF850 > $OUT1

# Add header
#head -140 $VCF850 > myanno_ALS_Cohort.vqsr_header.vcf #only need to do
    this once
cat myanno_ALS_Cohort.vqsr_header.vcf $OUT1 > $OUT2
sed '1,139d' $OUT2 > $OUT3
```

A.2.7 WGS cohort subsetting

This script was written by Ingrid Tarr and later modified by the candidate, and was used to subset the different cohorts from the 850-sample VCF after application of the Script A.2.6.

```
# WGS_gene_search_cohort_subsetting.R

# this code is for subsetting WGS gene search results into distinct cohorts for downstream analysis

# change project code MINE to SALS
# write script to get the list of IDs for whatever project code (or combo of codes) and use
# that to pull out the matching variant results, given a variant file

library(gdata)
setwd("/Volumes/data_FMHS/Restrict/Blair Group/Genetics/WGS_gene_searches/QC_and_analysis")

# this is the file that has the IDs and project code
full <- read.xls("/Volumes/data_FMHS/Restrict/Blair\ Group/Genetics/Project\ MiNE/Manifests/Master_manifest-850sequenced.xlsx", header = T, sheet = 1, stringsAsFactors = F, nrow = 1000)

# sample ID and Blair experiment code are the columns of interest
table(full$Blair.experiment.code)

SALS <- full[full$Blair.experiment.code == "SALS", "SampleID"]
FTD <- full[full$Blair.experiment.code == "FTD", "SampleID"]
FALS <- full[full$Blair.experiment.code == "FALS", "SampleID"]
SOD1 <- full[full$Blair.experiment.code == "SOD1", "SampleID"]
twin <- full[full$Blair.experiment.code == "Twin", "SampleID"]
twin_sod1 <- full[full$Blair.experiment.code == "Twin-SOD1", "SampleID"]

# change file path to variant file ## this needs to be updated each time
gene <- read.delim("/Volumes/data_FMHS/Restrict/Blair\ Group/Genetics/WGS_gene_searches/raw_data/HPC_resultant_txt/
```

```
                GENE_variants_headed.txt", header = T, skip = 0, sep =
                    "\t")

# if there are still some samples referred to as WIL..... then these two
    lines should replace them with the MQIDs
# if there aren't then it won't run these two lines
matchup <- if(length(grep("WIL", colnames(gene))) > 0){
   read.xls("/Volumes/data_FMHS/Restrict/Blair\ Group/Genetics/Project\
       MiNE/WIL\ ID\ conversion.xlsx", header = T, stringsAsFactors = F,
       col.names = c("tube", "mq", "wil", "manifest", "full_tube_id",
       "fastQ"))
   }
colnames(gene)[grep("WIL", colnames(gene))] <- if(length(grep("WIL",
    colnames(gene))) > 0){
   as.character(matchup[match(colnames(gene)[grep("WIL", colnames(gene))],
       matchup$wil), "mq"])
   }

# tidying of the MQIDs in the variant data for matching
colnames(gene)[grep("MQ160198", colnames(gene))] <- "12-MQ160198"
 colnames(gene) <- gsub("[[:punct:]]", "-", colnames(gene))
 colnames(gene) <- gsub("^X", "", colnames(gene))

# create a new data frame with the variant info for the group of interest:
#dat <- gene[, c(rep(T, 9), colnames(gene)[10:ncol(gene)] %in% SALS)]
SALS.gene <- gene[, c(rep(T, 9), colnames(gene)[10:ncol(gene)] %in% SALS)]
FTD.gene <- gene[, c(rep(T, 9), colnames(gene)[10:ncol(gene)] %in% FTD)]
# can also pull out combined data for groups as shown here:
# dat <- variant[, c(rep(T, 9), colnames(variant)[10:ncol(variant)] %in%
    c(SALS, FALS)]

# when saving, change the filepath
#write.csv(dat, "./dat.csv")
write.csv(SALS.gene, file = "SALS_gene_WGS_search.csv", row.names = FALSE)
write.csv(FTD.gene, file = "FTD_gene_WGS_search.csv", row.names = FALSE)
```

A.2.8 Novel nonsynonymous variant analysis of SALS WGS candidate gene screening results

This script was used to identify novel non-synonymous variants in a candidate gene among SALS patient WGS data, after application of the Scripts A.2.6 and A.2.7.

```
# WGS_gene_search_novel_nonsyn_analysis.R

# this code is for looking for novel variants in WGS gene search results in
    SALS and FTD patients

setwd("/Volumes/data_FMHS/Restrict/Blair
    Group/Genetics/WGS_gene_searches/QC_and_analysis/GENE")

library(stringr)
library(data.table)
library(WriteXLS)
library(readr)

# import data
SALS.gene <- read.csv("SALS_GENE_WGS_search.csv")

##### novel nonsynonymous variant analysis #####

# subset nonsynonymous
SALS.gene.nonsynonymous <-
    SALS.gene[grep("ExonicFunc.refGene=nonsynonymous", SALS.gene$INFO), ]

# Are any novel?
SALS.gene.nonsynonymous.novel <-
    SALS.gene.nonsynonymous[grep("ExAC_ALL=\\.",
    SALS.gene.nonsynonymous$INFO), ]
SALS.gene.nonsynonymous.novel <-
    SALS.gene.nonsynonymous.novel[grep("gnomAD_exome_ALL=\\.",
    SALS.gene.nonsynonymous.novel$INFO), ]
SALS.gene.nonsynonymous.novel <-
    SALS.gene.nonsynonymous.novel[grep("gnomAD_genome_ALL=\\.",
    SALS.gene.nonsynonymous.novel$INFO), ]
```

```
SALS.gene.nonsynonymous.novel <-
    SALS.gene.nonsynonymous.novel[grep("avsnp147=\\.",
    SALS.gene.nonsynonymous.novel$INFO), ]

# What are the variants?
x <- str_match(SALS.gene.nonsynonymous.novel$INFO,
    "AAChange.refGene=(.*?);")
x[,2] # this will print the AA change to the R console

# who are they in? ## ensure you do this for each line in your
    candidate.nonsynonymous.novel dataframe
which(apply(SALS.gene.nonsynonymous.novel[1,], 2, function(x)
    any(!grepl("0/0|\\./\\.", x))))
which(apply(SALS.gene.nonsynonymous.novel[2,], 2, function(x)
    any(!grepl("0/0|\\./\\.", x))))
```

A.2.9 Association analysis of SALS WGS candidate gene screening results

This script was used to identify any SNPs over or under represented among SALS patient WGS data compared with gnomAD and MGRB control individuals, after application of the scripts A.2.6 and A.2.7.

```
# WGS_gene_search_assoc_analysis.R

# this code is for looking for associated variants in WGS gene search
    results in SALS and FTD patients

setwd("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/QC_and_analysis/TIA1")

### data import ###

library(readr)

# import data
SALS.gene <- read_csv("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/QC_and_analysis/TIA1/SALS_gene_WGS_
    search.csv", col_types = cols(X1 = col_skip()))

# import allele counts from gnomAD # both means genomes and exomes
load("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/raw_data/gnomAD_data/gnomAD_allele_
    count_data_ALL_15-11-17.RObject")
load("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/raw_data/gnomAD_data/gnomAD_allele_
    count_data_NFE_15-11-17.RObject")

# import allele counts from welderly
load("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/raw_data/welderly_data/welderly_
    biallelic_small.Rdata")
load("/Volumes/data_FMHS/Restrict/Blair\
    Group/Genetics/WGS_gene_searches/raw_data/welderly_data/welderly_
    multiallelic_small.Rdata")

```

```
22
23 ### genotype and allele counting ###
24
25 # Count how many patients are hom, het and WT
26 SALS.gene$No.hom_patients <- apply(SALS.gene[,10:637], 1, function(u)
       length(which(grepl("1/1",u))==TRUE) )
27 SALS.gene$No.het_patients <- apply(SALS.gene[,10:637], 1, function(u)
       length(which(grepl("0/1|1/0",u))==TRUE) )
28 SALS.gene$No.WT_patients <- apply(SALS.gene[,10:637], 1, function(u)
       length(which(grepl("0/0",u))==TRUE) )
29
30 # Calculate the number of ref at alt alleles among the patiennts
31 SALS.gene$No.total_alleles <- (2*(SALS.gene$No.hom_patients) +
       2*(SALS.gene$No.het_patients) + 2*(SALS.gene$No.WT_patients))
32 SALS.gene$No.ref_alleles <- (1*(SALS.gene$No.het_patients) +
       2*(SALS.gene$No.WT_patients))
33 SALS.gene$No.alt_alleles <- (2*(SALS.gene$No.hom_patients) +
       1*(SALS.gene$No.het_patients))
34
35 # add exact.position column for merging
36 SALS.gene$exact.position <- paste(SALS.gene$`-CHROM`, SALS.gene$POS, sep =
       ":")
37
38
39 ### association testing with gnomAD ###
40
41 # appened gnomAD ALL and NFE (non-Finnish European) allele counts
42 SALS.gene.gnomAD <- merge(SALS.gene, gnomAD.allele.count.data.ALL.both, by
       = "exact.position", all.x = TRUE)
43 SALS.gene.gnomAD <- merge(SALS.gene.gnomAD,
       gnomAD.allele.count.data.NFE.both, by = "exact.position", all.x = TRUE)
44
45 # correct column classes
46 SALS.gene.gnomAD$gnomAD_both_ALL_total_allele_count <-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_ALL_total
       _allele_count))
47 SALS.gene.gnomAD$gnomAD_both_ALL_ref_allele_count <-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_ALL_ref
       _allele_count))
```

```
48 SALS.gene.gnomAD$gnomAD_both_ALL_alt_allele_count <-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_ALL_alt
       _allele_count))
49 SALS.gene.gnomAD$gnomAD_both_NFE_total_allele_count <-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_NFE_total
       _allele_count))
50 SALS.gene.gnomAD$gnomAD_both_NFE_ref_allele_count <-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_NFE_ref
       _allele_count))
51 SALS.gene.gnomAD$gnomAD_both_NFE_alt_allele_count<-
       as.numeric(as.character(SALS.gene.gnomAD$gnomAD_both_NFE_alt
       _allele_count))
52
53 # carry out fishers exact testing using allele counts from patients and ALL
       gnomAD controls to test for association
54 res <- NULL
55 for (i in 1:nrow(SALS.gene.gnomAD)){
56   table <- matrix(c(SALS.gene.gnomAD[i,643], SALS.gene.gnomAD[i,644],
         SALS.gene.gnomAD[i,646], SALS.gene.gnomAD[i,647]), ncol = 2, byrow =
         TRUE)
57   # if any NA occurs in your table save an error in p else run the fisher
         test
58   if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
59   # save all p values in a vector
60   res <- c(res,p)
61 }
62 SALS.gene.gnomAD$fishers.ALL <- res
63
64 # carry out fishers exact testing using allele counts from patients and NFE
       gnomAD controls to test for association
65 res <- NULL
66 for (i in 1:nrow(SALS.gene.gnomAD)){
67   table <- matrix(c(SALS.gene.gnomAD[i,643], SALS.gene.gnomAD[i,644],
         SALS.gene.gnomAD[i,649], SALS.gene.gnomAD[i,650]), ncol = 2, byrow =
         TRUE)
68   # if any NA occurs in your table save an error in p else run the fisher
         test
69   if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
70   # save all p values in a vector
```

```
  res <- c(res,p)
}
SALS.gene.gnomAD$fishers.NFE <- res


### association testing with welderly ###

# appened welderly allele counts
SALS.gene.welderly <- merge(SALS.gene, welderly_biallelic_small, by =
    "exact.position", all.x = TRUE)
SALS.gene.welderly <- merge(SALS.gene.welderly,
    welderly_multiallelic_small, by = "exact.position", all.x = TRUE)

# correct column classes
SALS.gene.welderly$AN_welderly <-
    as.numeric(as.character(SALS.gene.welderly$AN_welderly))
SALS.gene.welderly$AR_welderly <-
    as.numeric(as.character(SALS.gene.welderly$AR_welderly))
SALS.gene.welderly$AC_welderly <-
    as.numeric(as.character(SALS.gene.welderly$AC_welderly))

# check for multiallelic variants in welderly
# which rows? # what are there exact positions?
which(SALS.gene.welderly$welderly_multiallelic_flag == "multiallelic")
x <- SALS.gene.welderly[which(SALS.gene.welderly$welderly_multiallelic_flag
    == "multiallelic"), ]
x$exact.position

# carry out fishers exact testing using allele counts from patients and
    welderly controls to test for association
res <- NULL
for (i in 1:nrow(SALS.gene.welderly)){
  table <- matrix(c(SALS.gene.welderly[i,643], SALS.gene.welderly[i,644],
      SALS.gene.welderly[i,646], SALS.gene.welderly[i,647]), ncol = 2, byrow
      = TRUE)
  # if any NA occurs in your table save an error in p else run the fisher
      test
  if(any(is.na(table))) p <- "error" else p <- fisher.test(table)$p.value
  # save all p values in a vector
```

```
  res <- c(res,p)
}
SALS.gene.welderly$fishers.welderly <- res

# create smaler dataframes for manual analysis
SALS.gene.gnomAD.small <- cbind(SALS.gene.gnomAD[,1:10],
    SALS.gene.gnomAD[,639:652])
write.csv(SALS.gene.gnomAD.small, file = "SALS_TIA1_gnomAD_small.csv")

SALS.gene.welderly.small <- cbind(SALS.gene.welderly[,1:10],
    SALS.gene.welderly[,639:649])
write.csv(SALS.gene.welderly.small, file =
    "SALS_TIA1_welderly_results_small.csv")

SALS.gene.gnomAD.association <- cbind(SALS.gene.gnomAD[,1:10],
    SALS.gene.gnomAD[,639:650], SALS.gene.welderly[,645:647],
    SALS.gene.gnomAD[,651:652], SALS.gene.welderly[,648:649])
write.csv(SALS.gene.gnomAD.association, file =
    "SALS_TIA1_gnomAD_welderly_association_small.csv")
```

A.2.10 Creation of family WES VCFs

This script was written by Kelly Williams and later modified by the candidate, and was used to create a WES VCF for each family with multiple informative individuals present in the 137-sample WES VCF.

```
#!/bin/sh

#  split.files.all.families.sh

# generate family vcfs that exclude sites without a called a genotype and only include sites that have an alternate allele present
bcftools view -o ../QC\ and\ analysis/FALS15/FALS15.called.SNPs.vcf -Ov -c1 -U -s 15-A210,15-A211 Brisbane_MND.Kelly.hg19_multianno.vcf
bcftools view -o ../QC\ and\ analysis/FALS45/FALS45.called.SNPs.vcf -Ov -c1 -U -s 45-040247,45-A334 Brisbane_MND.Kelly.hg19_multianno.vcf
bcftools view -o ../QC\ and\ analysis/FALSmq2/FALSmq2.called.SNPs.vcf -Ov -c1 -U -s mq2-MQ140023,MQ130016 Brisbane_MND.Kelly.hg19_multianno.vcf
bcftools view -o ../QC\ and\ analysis/FALSmq20/FALSmq20.called.SNPs.vcf -Ov -c1 -U -s mq20-MQ140178,MQ130004 Brisbane_MND.Kelly.hg19_multianno.vcf

# create txt files with only column headers for ease of analysis
sed 's/#CHROM/CHROM/g' FALS15.called.SNPs.vcf > FALS15.called.SNPs.header.txt
sed 's/#CHROM/CHROM/g' FALS45.called.SNPs.vcf > FALS45.called.SNPs.header.txt
sed 's/#CHROM/CHROM/g' FALSmq2.called.SNPs.vcf > FALSmq2.called.SNPs.header.txt
sed 's/#CHROM/CHROM/g' FALSmq20.called.SNPs.vcf > FALSmq20.called.SNPs.header.txt
```

A.2.11 WES shared variant analysis for small families

This Rmarkdown code was co-written by the candidate with Kelly Williams and was used to identify a list of shared variants in each small family. It first identified all shared variants present among all affected members or a family and/or absent from any "married-in" control individuals, and then removed any shared variants that did not meet filtering criteria.

```
---
title: "small_families_exome_shared_variants.Rmd"
output: html_document
---

# This code identifies shared variants in each small ALS family and filters the resulting shared variants for population-based variants and non-protein-altering variants

# Set up
```{r setup}
# define working directory
directory <- "/Volumes/Research/MND/Ian Blair Group/Genetics/Exome relatedness PLINK/QC and analysis"

# generate an annotated RObject - ONLY DO ONCE
annot.vcf <- read.delim("/Volumes/Research/MND/Ian Blair Group/Genetics/Exome relatedness PLINK/Raw data/Brisbane_MND.Kelly.hg19_multianno.xls")
annot.vcf$exact.position <- paste(annot.vcf$Chr, annot.vcf$Start, sep = ":")
setwd(directory)
save(annot.vcf, file="Annotated_full_vcf.RObject")

```

#Load the family data into R
```{r load.all.families}

setwd(directory)

# note these have been generated using bcftools - "split.files.all.families.sh"
```
```

```
27
28 FALS15.called.SNPs <- read.table("/Volumes/Research/MND/Ian Blair
       Group/Genetics/Exome relatedness PLINK/QC and
       analysis/FALS15/FALS15.called.SNPs.header.txt", header=TRUE, quote="\"")
29 FALS45.called.SNPs <- read.table("/Volumes/Research/MND/Ian Blair
       Group/Genetics/Exome relatedness PLINK/QC and
       analysis/FALS45/FALS45.called.SNPs.header.txt", header=TRUE, quote="\"")
30 FALSmq2.called.SNPs <- read.table("/Volumes/Research/MND/Ian Blair
       Group/Genetics/Exome relatedness PLINK/QC and
       analysis/FALSmq2/FALSmq2.called.SNPs.header.txt", header=TRUE,
       quote="\"")
31 FALSmq20.called.SNPs <- read.table("/Volumes/Research/MND/Ian Blair
       Group/Genetics/Exome relatedness PLINK/QC and
       analysis/FALSmq20/FALSmq20.called.SNPs.header.txt", header=TRUE,
       quote="\"")
32
33 load("Annotated_full_vcf.RObject")
34 annot.info <- annot.vcf[,c(1:58,205)]
35
36 ```
37
38 ##FALS15 analysis##
39 ```{r affected.only.fals15}
40
41 setwd(directory)
42
43 # retain SNPs that are present in both individuals (as homozygous or
       heterozygous)
44 FALS15.shared.SNPs <-
       FALS15.called.SNPs[Reduce('&',lapply(FALS15.called.SNPs[10:11],
       function(x) grepl("0/1|1/0|1/1", x))),]
45
46 # generate a location column for merging purposes
47 FALS15.shared.SNPs$exact.position <- paste(FALS15.shared.SNPs$CHROM,
       FALS15.shared.SNPs$POS, sep = ":")
48
49 # merge the files so you have annotated variants for filtering
50 # note that you really only want to merge the "info" columns
```

```
FALS15.annotated.shared.SNPs <- merge(FALS15.shared.SNPs, annot.info,
    by="exact.position", all.x=TRUE)

### filtering shared variants ###
# perform filtering steps to get final number of novel shared exonic SNPs
# keep only variants with an annotation
filter1 <- which(FALS15.annotated.shared.SNPs$snp129 == ".") # remove
    dbSNP129
filtered.SNPs <- FALS15.annotated.shared.SNPs[filter1,]
filter2 <- which(filtered.SNPs$X1000g2014oct_all == ".") # remove 1000
    Genomes
filtered.SNPs <- filtered.SNPs[filter2,]
filter3 <- which(filtered.SNPs$ExonicFunc.refGene != ".") # remove
    non-coding
filtered.SNPs <- filtered.SNPs[filter3,]
filter4 <- which(filtered.SNPs$avsnp142 == ".") # remove dbSNP142
filtered.SNPs <- filtered.SNPs[filter4,]
filter5 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV") #
    remove synonymous
filtered.SNPs <- filtered.SNPs[filter5,]
FALS15.filtered.novel.shared.SNPs <- filtered.SNPs # shared variants

# export to csv
write.table(FALS15.filtered.novel.shared.SNPs, "FALS15.variants.txt", eol =
    "\r", quote=FALSE, sep="\t", row.names=FALSE)

```

```
##FALS45 analysis##
```{r affected.only.fals45}

setwd(directory)

# retain SNPs that are present in all 4 individuals (as homozygous or
    heterozygous)
FALS45.shared.SNPs <-
    FALS45.called.SNPs[Reduce('&',lapply(FALS45.called.SNPs[12:12],
    function(x) grepl("0/1|1/0|1/1", x))),]

```
```

```
# remove SNPs that are present in the control sample
control <- which(colnames(FALS45.shared.SNPs) == "X45.040542")
FALS45.SNPs.no.control <-
    FALS45.shared.SNPs[Reduce('|',lapply(FALS45.shared.SNPs[control],
    function(x) grepl("\\.\\/\\.|0/0", x))),]

# generate a location column for merging purposes
FALS45.SNPs.no.control$exact.position <-
    paste(FALS45.SNPs.no.control$CHROM, FALS45.SNPs.no.control$POS, sep =
    ":")

# merge the files so you have annotated variants for filtering
# note that you really only want to merge the "info" columns
FALS45.annotated.shared.SNPs <- merge(FALS45.SNPs.no.control, annot.info,
    by="exact.position", all.x=TRUE)


### filtering shared variants ###
# perform filtering steps to get final number of novel shared exonic SNPs
# keep only variants with an annotation
filter1 <- which(FALS45.annotated.shared.SNPs$snp129 == ".") # remove
    dbSNP129
filtered.SNPs <- FALS45.annotated.shared.SNPs[filter1,]
filter2 <- which(filtered.SNPs$X1000g2014oct_all == ".") # remove 1000
    Genomes
filtered.SNPs <- filtered.SNPs[filter2,]
filter3 <- which(filtered.SNPs$ExonicFunc.refGene != ".") # remove
    non-coding
filtered.SNPs <- filtered.SNPs[filter3,]
filter4 <- which(filtered.SNPs$avsnp142 == ".") # remove dbSNP142
filtered.SNPs <- filtered.SNPs[filter4,]
filter5 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV") #
    remove synonymous
filtered.SNPs <- filtered.SNPs[filter5,]
FALS15.filtered.novel.shared.SNPs <- filtered.SNPs # shared variants

#export to csv
write.table(FALS45.filtered.novel.shared.SNPs, "FALS45.variants.txt", eol =
    "\r", quote=FALSE, sep="\t", row.names=FALSE)
```

```
110
111 ```
112
113 ##FALSmq2 analysis##
114 ```{r affected.only.falsmq2}
115
116 setwd(directory)
117
118 # retain SNPs that are present in both individuals (as homozygous or
    heterozygous)
119 FALSmq2.shared.SNPs <-
    FALSmq2.called.SNPs[Reduce('&',lapply(FALSmq2.called.SNPs[10:11],
    function(x) grepl("0/1|1/0|1/1", x))),]
120
121 # generate a location column for merging purposes
122 FALSmq2.shared.SNPs$exact.position <- paste(FALSmq2.shared.SNPs$CHROM,
    FALSmq2.shared.SNPs$POS, sep = ":")
123
124 # merge the files so you have annotated variants for filtering
125 # note that you really only want to merge the "info" columns
126 FALSmq2.annotated.shared.SNPs <- merge(FALSmq2.shared.SNPs, annot.info,
    by="exact.position", all.x=TRUE)
127
128 ### filtering shared variants ###
129 # perform filtering steps to get final number of novel shared exonic SNPs
130 # keep only variants with an annotation
131 filter1 <- which(FALSmq2.annotated.shared.SNPs$snp129 == ".") # remove
    dbSNP129
132 filtered.SNPs <- FALSmq2.annotated.shared.SNPs[filter1,]
133 filter2 <- which(filtered.SNPs$X1000g2014oct_all == ".") # remove 1000
    Genomes
134 filtered.SNPs <- filtered.SNPs[filter2,]
135 filter3 <- which(filtered.SNPs$ExonicFunc.refGene != ".") # remove
    non-coding
136 filtered.SNPs <- filtered.SNPs[filter3,]
137 filter4 <- which(filtered.SNPs$avsnp142 == ".") # remove dbSNP142
138 filtered.SNPs <- filtered.SNPs[filter4,]
139 filter5 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV") #
    remove synonymous
```

```
filtered.SNPs <- filtered.SNPs[filter5,]
FALS15.filtered.novel.shared.SNPs <- filtered.SNPs # shared variants

#export to csv
write.table(FALSmq2.filtered.novel.shared.SNPs, "FALSmq2.variants.txt", eol = "\r", quote=FALSE, sep="\t", row.names=FALSE)

```

```
##FALSmq20 analysis##
```{r affected.only.falsmq20}

setwd(directory)

# retain SNPs that are present in all 4 individuals (as homozygous or heterozygous)
FALSmq20.shared.SNPs <- FALSmq20.called.SNPs[Reduce(`&`,lapply(FALSmq20.called.SNPs[10:11], function(x) grepl("0/1|1/0|1/1", x))),]

# generate a location column for merging purposes
FALSmq20.shared.SNPs$exact.position <- paste(FALSmq20.shared.SNPs$CHROM, FALSmq20.shared.SNPs$POS, sep = ":")

# merge the files so you have annotated variants for filtering
# note that you really only want to merge the "info" columns
FALSmq20.annotated.shared.SNPs <- merge(FALSmq20.shared.SNPs, annot.info, by="exact.position", all.x=TRUE)

### filtering shared variants ###
# perform filtering steps to get final number of novel shared exonic SNPs
# keep only variants with an annotation
filter1 <- which(FALSmq20.annotated.shared.SNPs$snp129 == ".") # remove dbSNP129
filtered.SNPs <- FALSmq20.annotated.shared.SNPs[filter1,]
filter2 <- which(filtered.SNPs$X1000g2014oct_all == ".") # remove 1000 Genomes
filtered.SNPs <- filtered.SNPs[filter2,]
```
```

```
filter3 <- which(filtered.SNPs$ExonicFunc.refGene != ".") # remove
    non-coding
filtered.SNPs <- filtered.SNPs[filter3,]
filter4 <- which(filtered.SNPs$avsnp142 == ".") # remove dbSNP142
filtered.SNPs <- filtered.SNPs[filter4,]
filter5 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV") #
    remove synonymous
filtered.SNPs <- filtered.SNPs[filter5,]
FALS15.filtered.novel.shared.SNPs <- filtered.SNPs # shared variants

#export to csv
write.table(FALSmq20.filtered.novel.shared.SNPs, "FALSmq20.variants.txt",
    eol = "\r", quote=FALSE, sep="\t", row.names=FALSE)

```

A.2.12 Combining WES VCFs for family FALSmq28

This script was used to combine the three single-sample WES VCFs for the informative family members from FALSmq28.

```
#!/bin/sh

#  mq28_combine_vcfs.sh

# this code is for creating a combined vcf for the 3 individauls from
    FALSmq28 who were exome sequenced at Macrogen

# line count each individual file
cd /Users/emccann/Desktop/FALSmq28_exomes/Raw\ data
wc -l ./mq28-MQ150214/mq28-MQ150214.final.vcf
wc -l ./mq28-MQ150267/mq28-MQ150267.final.vcf
wc -l ./mq28-MQ150303/mq28-MQ150303.final.vcf

# bgzip these vcf files so we can merge them with bcftools
cd /Users/emccann/Desktop/FALSmq28_exomes/QC\ and\ analysis
bgzip -c ../Raw\ data/mq28-MQ150214/mq28-MQ150214.final.vcf >
    mq28-MQ150214.final.vcf.gz
bgzip -c ../Raw\ data/mq28-MQ150267/mq28-MQ150267.final.vcf >
    mq28-MQ150267.final.vcf.gz
bgzip -c ../Raw\ data/mq28-MQ150303/mq28-MQ150303.final.vcf >
    mq28-MQ150303.final.vcf.gz

# index these vcf files so we can merge them with bcftools
tabix -p vcf mq28-MQ150214.final.vcf.gz
tabix -p vcf mq28-MQ150267.final.vcf.gz
tabix -p vcf mq28-MQ150303.final.vcf.gz

# merge individual vcf to create a combined vcf file
cd /Users/emccann/Desktop/FALSmq28_exomes/QC\ and\ analysis
bcftools merge mq28-MQ150214.final.vcf.gz mq28-MQ150267.final.vcf.gz
    mq28-MQ150303.final.vcf.gz > FALSmq28.final.vcf.gz

# decompress the combined vcf file
bgzip -d FALSmq28.final.vcf.gz

```

```
# generate a family vcf that excludes sites without a called a genotype
bcftools view -o FALSmq28.final.called.vcf -Ov -U -s
    mq28-MQ150214,mq28-MQ150267,mq28-MQ150303 FALSmq28.final.vcf

# generate vcfs that only have an alternate allele present
bcftools view -o FALSmq28.final.called.SNPs.vcf -Ov -c1 -U -s
    mq28-MQ150214,mq28-MQ150267,mq28-MQ150303 FALSmq28.final.called.vcf
```

A.2.13 WES shared variant analysis for family FALSmq28

This script was used to identify a list of shared variants in FALSmq28. It first identified all shared variants present among both affected family members and the obligate mutation carrier, and then removed any shared variants that did not meet filtering criteria.

```
# mq28_exome_shared_variants.R

# This script is for shared variant analysis of exome sequencing data from FALSmq28

# first import the tab delimited ANNOVAR annotated combined VCF
mq28 <- read.csv("/Users/emccann/Desktop/FALSmq28_exomes/QC\ and\ analysis/FALSmq28exomes_anno.hg19_multianno_headed.csv", header = TRUE, sep = ",") # 185 703 lines

# set working directory
setwd("~/Desktop/FALSmq28_exomes/QC and analysis")

# Start shared variant analysis
## note: FALSmq28 has 2 affected patients and an obligate carrier - there are no married in controls

# retain SNPs that are present in all 4 individuals (as homozygous or heterozygous)
mq28.shared.SNPs.alt1 <- mq28[Reduce('&',lapply(mq28[126:128], function(x) grepl("0/1|1/0|1/1|1/2|2/1|1/3|3/1|1/4|4/1|\\.\\/\\.", x))),] # 182 851 lines
mq28.shared.SNPs.alt2 <- mq28[Reduce('&',lapply(mq28[126:128], function(x) grepl("0/2|2/0|2/2|2/1|1/2|2/3|3/2|2/4|4/2|\\.\\/\\.", x))),] # 388 lines
mq28.shared.SNPs.alt3 <- mq28[Reduce('&',lapply(mq28[126:128], function(x) grepl("0/3|3/0|3/3|3/1|1/3|3/2|2/3|3/4|4/3|\\.\\/\\.", x))),] # 0 lines
mq28.shared.SNPs.alt4 <- mq28[Reduce('&',lapply(mq28[126:128], function(x) grepl("0/4|4/0|4/4|4/1|1/4|4/2|2/4|4/3|3/4|\\.\\/\\.", x))),] # 0 lines
mq28.shared.SNPs.ALL <- rbind(mq28.shared.SNPs.alt1, mq28.shared.SNPs.alt2, mq28.shared.SNPs.alt3, mq28.shared.SNPs.alt4)

# generate a location column for merging purposes
```

```
mq28.shared.SNPs.ALL$exact.position <- paste(mq28.shared.SNPs.ALL$CHROM,
    mq28.shared.SNPs.ALL$POS, sep = ":")
# get rid of the "chr"
mq28.shared.SNPs.ALL$exact.position <- sub("chr", "",
    mq28.shared.SNPs.ALL$exact.position , perl=T)

### filtering shared variants ###
# perform filtering steps to get final number of novel shared exonic SNPs
# keep only variants with an annotation
filter1 <- which(mq28.shared.SNPs.ALL$avsnp147 == ".") # remove dbSNP147
filtered.SNPs <- mq28.shared.SNPs.ALL[filter1,] # 3 378 lines
# remove variants with no exonic function
filter2 <- which(filtered.SNPs$Func.refGene == "exonic")
filtered.SNPs <- filtered.SNPs[filter2,] # 230 variants
# remove synonymous variants
filter3 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV")
filtered.SNPs <- filtered.SNPs[filter3,] # 171 variants

# merge remaining variants with gnomAD.vcf.data for further filtering in
    excel
load("/Users/emccann/Desktop/FALSmq28_exomes/Raw\
    data/gnomAD_genomes.vcf.data.RObject")
load("/Users/emccann/Desktop/FALSmq28_exomes/Raw\
    data/gnomAD_exomes_vcf_data.RObject")
filtered.SNPs <- merge(filtered.SNPs, gnomAD.genomes.vcf.data, by =
    "exact.position", all.x = TRUE)
filtered.SNPs <- merge(filtered.SNPs, gnomAD.exomes.vcf.data, by =
    "exact.position", all.x = TRUE)

x <- filtered.SNPs

# export to csv to use in excel
write.table(x, "FALSmq28_exome_shared_variants.txt", quote=FALSE, sep="\t",
    row.names=FALSE, eol = "\r")
```

A.2.14 WGS shared variant analysis for family FALSmq28

This script was used to identify a list of shared variants in FALSmq28. It first identified all shared variants present among both affected family members and the obligate mutation carrier, and then removed any shared variants that did not meet filtering criteria.

````
---
title: "mq28_exome_shared_variants.Rmd"
output: html_document
---

# This code identifies shared variants in FALSmq28 and filters the resulting shared variants for population-based variants and non-protein-altering variants

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Set the working directory

```{r directories}
setwd("~/Desktop/WGS_shared_variant_analysis/QC and analysis")
```

###### Analysis 1 ######

## Load family data into R
## annotated family WES or WGS VCF - either complete (analysis 1), containing regions with LOD>0 (analysis 2) or
## containing regions with LOD>-2 and variants with GQ>20 (analysis 3)
```{r load.vcf}
FALSmq28 <- read.delim("~/Desktop/WGS_shared_variant_analysis/Raw\ data/Filter_first/FALSmq28_anno.hg19_multianno_headed.txt", header = TRUE, sep = "\t")
```

## Shared variant analysis and first-tier filtering
```{r sharedVariants}
````
````

```
29
30 # generate a location column for merging purposes
31 FALSmq28$exact.position <- paste(FALSmq28$CHROM, FALSmq28$POS, sep = ":")
32
33 # retain SNPs that are present in all 3 individuals (as homozygous or
       heterozygous)
34 FALSmq28.shared.SNPs.alt1 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/1|1/0|1/1|1/2|2/1|1/3|3/1|1/4|4/1|1/5|5/1|1/6|6/1|\\.\\/\\.",
       x))),] # 3361529 variants
35 FALSmq28.shared.SNPs.alt2 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/2|2/0|2/2|2/1|1/2|2/3|3/2|2/4|4/2|2/5|5/2|2/6|6/2|\\.\\/\\.",
       x))),] # 251193 variants
36 FALSmq28.shared.SNPs.alt3 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/3|3/0|3/3|3/1|1/3|3/2|2/3|3/4|4/3|3/5|5/3|3/6|6/3|\\.\\/\\.",
       x))),] # 66939 variants
37 FALSmq28.shared.SNPs.alt4 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/4|4/0|4/4|4/1|1/4|4/2|2/4|4/3|3/4|4/5|5/4|4/6|6/4|\\.\\/\\.",
       x))),] # 24723 variants
38 FALSmq28.shared.SNPs.alt5 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/5|5/0|5/5|5/1|1/5|5/2|2/5|5/3|3/5|5/4|4/5|5/6|6/5|\\.\\/\\.",
       x))),] # 9841 variants
39 FALSmq28.shared.SNPs.alt6 <- FALSmq28[Reduce('&',lapply(FALSmq28[126:128],
       function(x)
       grepl("0/6|6/0|6/6|6/1|1/6|6/2|2/6|6/3|3/6|6/4|4/6|6/5|5/6|\\.\\/\\.",
       x))),] # 4704 variants
40 FALSmq28.shared.SNPs.ALL <- rbind(FALSmq28.shared.SNPs.alt1,
       FALSmq28.shared.SNPs.alt2, FALSmq28.shared.SNPs.alt3,
       FALSmq28.shared.SNPs.alt4, FALSmq28.shared.SNPs.alt5,
       FALSmq28.shared.SNPs.alt6) # 3718299 annotations
41
42 length(unique(FALSmq28.shared.SNPs.ALL$exact.position)) # 2792679 variants
43
44 ### filtering shared variants ###
45 # perform filtering steps to get final number of novel shared SNPs
```

```
# different variations and combinations of these steps were applied in the different analysis pipelines
# remove known SNPs from dbSNP147
filter1 <- which(FALSmq28.shared.SNPs.ALL$avsnp147 == ".") # remove dbSNP147
filtered.SNPs <- FALSmq28.shared.SNPs.ALL[filter1,]
length(unique(filtered.SNPs$exact.position)) # number of unique variants
# remove variants with no exonic function
filter2 <- which(filtered.SNPs$Func.refGene == "exonic")
filtered.SNPs <- filtered.SNPs[filter2,]
length(unique(filtered.SNPs$exact.position)) # number of unique variants
# remove synonymous variants
filter3 <- which(filtered.SNPs$ExonicFunc.refGene != "synonymous SNV")
filtered.SNPs <- filtered.SNPs[filter3,]
length(unique(filtered.SNPs$exact.position)) # number of unique variants
# remove intronic and intergenic variants
filter4 <- which(filtered.exome.SNPs$Func.refGene != "intronic")
filtered.exome.SNPs <- filtered.exome.SNPs[filter4,]
filter5 <- which(filtered.exome.SNPs$Func.refGene != "intergenic")
filtered.exome.SNPs <- filtered.exome.SNPs[filter5,]
filter6 <- which(filtered.exome.SNPs$Func.refGene != "ncRNA_intronic")
filtered.exome.SNPs <- filtered.exome.SNPs[filter6,]

# merge remaining variants with gnomAD.vcf.data for further filtering in excel
load("/Users/emilymccann/Desktop/WGS_shared_variant_analysis/QC\ and\ analysis/gnomAD.genomes.vcf.data.RObject")
load("/Users/emilymccann/Desktop/WGS_shared_variant_analysis/QC\ and\ analysis/gnomAD_exomes_vcf_data.RObject")
filtered.SNPs <- merge(filtered.SNPs, gnomAD.genomes.vcf.data, by = "exact.position", all.x = TRUE)
filtered.SNPs <- merge(filtered.SNPs, gnomAD.exomes.vcf.data, by = "exact.position", all.x = TRUE)

x <- filtered.SNPs

# export to csv to use in excel
write.table(x, "FALSmq28_shared_variants.txt", quote=FALSE, sep="\t", row.names=FALSE, eol = "\r")
```

A.2.15 File preparation for genetic linkage analysis of FALSmq28 in Merlin

This script was used to edit ped, dat and map files for FALSmq28 to prepare these for genetic linkage analysis using Merlin.

```
# mq28_setup_linkage.R

# this code is for carrying out linkage analysis of FALSmq28

# First install the paramlink package
install.packages("paramlink")

# Now load the paramlink package
library("paramlink")

# Set working directory
setwd("/Volumes/Personal/Bioinformatics/Linkage/QC and analysis")

# Load data
mq28.ped <- read.table("/Volumes/Personal/Bioinformatics/Linkage/Raw\
    data/mq28/test.ped")
mq28.merlin.ped <- read.table("/Volumes/GEORGE/merlin-1.1.2\
    copy/mq28_raw_files/mq28_w_liability.ped")

# Make mq28 ped into linkdat object
x <- linkdat(mq28.ped)

# let's have a look at the pedogree
plot(x, available=TRUE)

# so we can see a few errors... the obligates are marked as unaffected and
    the "unknown" is not who it should be...
## We need to fix the "aff" column 5 to reflect the proper statuses of
    these individuals
mq28.ped[1, 6] <- 1
mq28.ped[2, 6] <- 2
mq28.ped[3, 6] <- 1
mq28.ped[4, 6] <- 2
mq28.ped[5, 6] <- 2
```

```
mq28.ped[6, 6] <- 1
mq28.ped[7, 6] <- 2
mq28.ped[8, 6] <- 0
mq28.ped[9, 6] <- 0
mq28.ped[10, 6] <- 0
mq28.ped[11, 6] <- 1
mq28.ped[12, 6] <- 0
mq28.ped[13, 6] <- 0
mq28.ped[14, 6] <- 2
mq28.ped[15, 6] <- 1
mq28.ped[16, 6] <- 0
mq28.ped[17, 6] <- 0
mq28.ped[18, 6] <- 2
mq28.ped[19, 6] <- 1
mq28.ped[20, 6] <- 2
mq28.ped[21, 6] <- 1
mq28.ped[22, 6] <- 0
mq28.ped[23, 6] <- 0
mq28.ped[24, 6] <- 0
mq28.ped[25, 6] <- 0

## Make mq28.ped back into a linkdat object
x = linkdat(mq28.ped)

## let's look at the pedigree again to see if it looks right now
plot(x, available=TRUE )
summary(x)
### Excellent! It is now correct

## Now, let's add the liability classes
mq28.ped$liability_class <- c(0, 1, 0, 6, 1, 0, 5, 6, 6, 6, 0, 3, 4, 4, 0,
    1, 2, 6, 0, 4, 0, 4, 4, 4, 4)
x2 = as.data.frame(mq28.ped)
write.table(x2, file = "mq28_w_liability_new.txt", sep = "\t", col.names =
    FALSE, row.names = FALSE)

# save this work as files we can use in merlin
write.linkdat(x, prefix="mq28", what=c("ped", "map", "dat", "freq",
    "model"), merlin=TRUE)
```

A.2.16 Splitting FALSmq28 genetic linkage analysis files by chromosome

This script was used to split the files created in section A.2.15 by chromosome, to facilitate running genetic linkage analysis using Merlin for each chromosome separately.

```
# mq28_split_linkage_files.R

# this code leads on from file_preparation_for_imputing_genotypes.R and is
    for the purposes of making ped map and dat files to conduct linkage
    analysis in merlin separately for each chr

setwd("/Volumes/Personal/Bioinformatics/Linkage/QC and
    analysis/Linkage_by_chr")

# combine chr ped files with libability class column
FALSmq28_chr1.ped2 <- cbind(FALSmq28_chr1.ped, ped[,79895])
FALSmq28_chr2.ped2 <- cbind(FALSmq28_chr2.ped, ped[,79895])
FALSmq28_chr3.ped2 <- cbind(FALSmq28_chr3.ped, ped[,79895])
FALSmq28_chr4.ped2 <- cbind(FALSmq28_chr4.ped, ped[,79895])
FALSmq28_chr5.ped2 <- cbind(FALSmq28_chr5.ped, ped[,79895])
FALSmq28_chr6.ped2 <- cbind(FALSmq28_chr6.ped, ped[,79895])
FALSmq28_chr7.ped2 <- cbind(FALSmq28_chr7.ped, ped[,79895])
FALSmq28_chr8.ped2 <- cbind(FALSmq28_chr8.ped, ped[,79895])
FALSmq28_chr9.ped2 <- cbind(FALSmq28_chr9.ped, ped[,79895])
FALSmq28_chr10.ped2 <- cbind(FALSmq28_chr10.ped, ped[,79895])
FALSmq28_chr11.ped2 <- cbind(FALSmq28_chr11.ped, ped[,79895])
FALSmq28_chr12.ped2 <- cbind(FALSmq28_chr12.ped, ped[,79895])
FALSmq28_chr13.ped2 <- cbind(FALSmq28_chr13.ped, ped[,79895])
FALSmq28_chr14.ped2 <- cbind(FALSmq28_chr14.ped, ped[,79895])
FALSmq28_chr15.ped2 <- cbind(FALSmq28_chr15.ped, ped[,79895])
FALSmq28_chr16.ped2 <- cbind(FALSmq28_chr16.ped, ped[,79895])
FALSmq28_chr17.ped2 <- cbind(FALSmq28_chr17.ped, ped[,79895])
FALSmq28_chr18.ped2 <- cbind(FALSmq28_chr18.ped, ped[,79895])
FALSmq28_chr19.ped2 <- cbind(FALSmq28_chr19.ped, ped[,79895])
FALSmq28_chr20.ped2 <- cbind(FALSmq28_chr20.ped, ped[,79895])
FALSmq28_chr21.ped2 <- cbind(FALSmq28_chr21.ped, ped[,79895])
FALSmq28_chr22.ped2 <- cbind(FALSmq28_chr22.ped, ped[,79895])
FALSmq28_chr23.ped2 <- cbind(FALSmq28_chr23.ped, ped[,79895])

```

```
32 # Write these to file
33 write.table(FALSmq28_chr1.ped2, file = "FALSmq28_chr1.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
34 write.table(FALSmq28_chr2.ped2, file = "FALSmq28_chr2.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
35 write.table(FALSmq28_chr3.ped2, file = "FALSmq28_chr3.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
36 write.table(FALSmq28_chr4.ped2, file = "FALSmq28_chr4.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
37 write.table(FALSmq28_chr5.ped2, file = "FALSmq28_chr5.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
38 write.table(FALSmq28_chr6.ped2, file = "FALSmq28_chr6.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
39 write.table(FALSmq28_chr7.ped2, file = "FALSmq28_chr7.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
40 write.table(FALSmq28_chr8.ped2, file = "FALSmq28_chr8.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
41 write.table(FALSmq28_chr9.ped2, file = "FALSmq28_chr9.ped.txt", sep = "\t",
       col.names = FALSE, row.names = FALSE)
42 write.table(FALSmq28_chr10.ped2, file = "FALSmq28_chr10.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
43 write.table(FALSmq28_chr11.ped2, file = "FALSmq28_chr11.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
44 write.table(FALSmq28_chr12.ped2, file = "FALSmq28_chr12.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
45 write.table(FALSmq28_chr13.ped2, file = "FALSmq28_chr13.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
46 write.table(FALSmq28_chr14.ped2, file = "FALSmq28_chr14.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
47 write.table(FALSmq28_chr15.ped2, file = "FALSmq28_chr15.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
48 write.table(FALSmq28_chr16.ped2, file = "FALSmq28_chr16.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
49 write.table(FALSmq28_chr17.ped2, file = "FALSmq28_chr17.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
50 write.table(FALSmq28_chr18.ped2, file = "FALSmq28_chr18.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
51 write.table(FALSmq28_chr19.ped2, file = "FALSmq28_chr19.ped.txt", sep =
       "\t", col.names = FALSE, row.names = FALSE)
```

```
write.table(FALSmq28_chr20.ped2, file = "FALSmq28_chr20.ped.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr21.ped2, file = "FALSmq28_chr21.ped.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr22.ped2, file = "FALSmq28_chr22.ped.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr23.ped2, file = "FALSmq28_chr23.ped.txt", sep = "\t", col.names = FALSE, row.names = FALSE)

# redo the simple version of the map files
map <- read.table("/Users/emccann/Desktop/merlin-1.1.2/mq28_raw_files/test.map")

#subset map file by chromosome
FALSmq28_chr1.map <- map[which(map$V1 == "1") ,]
FALSmq28_chr2.map <- map[which(map$V1 == "2") ,]
FALSmq28_chr3.map <- map[which(map$V1 == "3") ,]
FALSmq28_chr4.map <- map[which(map$V1 == "4") ,]
FALSmq28_chr5.map <- map[which(map$V1 == "5") ,]
FALSmq28_chr6.map <- map[which(map$V1 == "6") ,]
FALSmq28_chr7.map <- map[which(map$V1 == "7") ,]
FALSmq28_chr8.map <- map[which(map$V1 == "8") ,]
FALSmq28_chr9.map <- map[which(map$V1 == "9") ,]
FALSmq28_chr10.map <- map[which(map$V1 == "10") ,]
FALSmq28_chr11.map <- map[which(map$V1 == "11") ,]
FALSmq28_chr12.map <- map[which(map$V1 == "12") ,]
FALSmq28_chr13.map <- map[which(map$V1 == "13") ,]
FALSmq28_chr14.map <- map[which(map$V1 == "14") ,]
FALSmq28_chr15.map <- map[which(map$V1 == "15") ,]
FALSmq28_chr16.map <- map[which(map$V1 == "16") ,]
FALSmq28_chr17.map <- map[which(map$V1 == "17") ,]
FALSmq28_chr18.map <- map[which(map$V1 == "18") ,]
FALSmq28_chr19.map <- map[which(map$V1 == "19") ,]
FALSmq28_chr20.map <- map[which(map$V1 == "20") ,]
FALSmq28_chr21.map <- map[which(map$V1 == "21") ,]
FALSmq28_chr22.map <- map[which(map$V1 == "22") ,]
FALSmq28_chr23.map <- map[which(map$V1 == "23") ,]

# write these to file
```

```
write.table(FALSmq28_chr1.map, file = "FALSmq28_chr1.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr2.map, file = "FALSmq28_chr2.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr3.map, file = "FALSmq28_chr3.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr4.map, file = "FALSmq28_chr4.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr5.map, file = "FALSmq28_chr5.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr6.map, file = "FALSmq28_chr6.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr7.map, file = "FALSmq28_chr7.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr8.map, file = "FALSmq28_chr8.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr9.map, file = "FALSmq28_chr9.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr10.map, file = "FALSmq28_chr10.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr11.map, file = "FALSmq28_chr11.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr12.map, file = "FALSmq28_chr12.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr13.map, file = "FALSmq28_chr13.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr14.map, file = "FALSmq28_chr14.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr15.map, file = "FALSmq28_chr15.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr16.map, file = "FALSmq28_chr16.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr17.map, file = "FALSmq28_chr17.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr18.map, file = "FALSmq28_chr18.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr19.map, file = "FALSmq28_chr19.map.txt", sep = "\t", col.names = FALSE, row.names = FALSE)
```

```
105 write.table(FALSmq28_chr20.map, file = "FALSmq28_chr20.map.txt", sep =
        "\t", col.names = FALSE, row.names = FALSE)
106 write.table(FALSmq28_chr21.map, file = "FALSmq28_chr21.map.txt", sep =
        "\t", col.names = FALSE, row.names = FALSE)
107 write.table(FALSmq28_chr22.map, file = "FALSmq28_chr22.map.txt", sep =
        "\t", col.names = FALSE, row.names = FALSE)
108 write.table(FALSmq28_chr23.map, file = "FALSmq28_chr23.map.txt", sep =
        "\t", col.names = FALSE, row.names = FALSE)
109
110 # redo do dat files
111 dat <-
        read.table("/Users/emccann/Desktop/merlin-1.1.2/mq28_raw_files/test.dat")
112
113 # using number of markers belonging to each chromosome worked out above -
        "# how many markers belong to each chromosome?"
114 # subset dat file based on these numbers
115 FALSmq28_chr1.dat <- rbind(dat[1, ], dat[2:3158, ], dat[39946, ],
        dat[39947, ])
116 FALSmq28_chr2.dat <- rbind(dat[1, ], dat[3159:6271, ], dat[39946, ],
        dat[39947, ])
117 FALSmq28_chr3.dat <- rbind(dat[1, ], dat[6272:8846, ], dat[39946, ],
        dat[39947, ])
118 FALSmq28_chr4.dat <- rbind(dat[1, ], dat[8847:11328, ], dat[39946, ],
        dat[39947, ])
119 FALSmq28_chr5.dat <- rbind(dat[1, ], dat[11329:13651, ], dat[39946, ],
        dat[39947, ])
120 FALSmq28_chr6.dat <- rbind(dat[1, ], dat[13652:16285, ], dat[39946, ],
        dat[39947, ])
121 FALSmq28_chr7.dat <- rbind(dat[1, ], dat[16286:18529, ], dat[39946, ],
        dat[39947, ])
122 FALSmq28_chr8.dat <- rbind(dat[1, ], dat[18530:20531, ], dat[39946, ],
        dat[39947, ])
123 FALSmq28_chr9.dat <- rbind(dat[1, ], dat[20532:22344, ], dat[39946, ],
        dat[39947, ])
124 FALSmq28_chr10.dat <- rbind(dat[1, ], dat[22345:24336, ], dat[39946, ],
        dat[39947, ])
125 FALSmq28_chr11.dat <- rbind(dat[1, ], dat[24337:26204, ], dat[39946, ],
        dat[39947, ])
```

```
126 FALSmq28_chr12.dat <- rbind(dat[1, ], dat[26205:28087, ], dat[39946, ],
        dat[39947, ])
127 FALSmq28_chr13.dat <- rbind(dat[1, ], dat[28088:29483, ], dat[39946, ],
        dat[39947, ])
128 FALSmq28_chr14.dat <- rbind(dat[1, ], dat[29484:30738, ], dat[39946, ],
        dat[39947, ])
129 FALSmq28_chr15.dat <- rbind(dat[1, ], dat[30739:31871, ], dat[39946, ],
        dat[39947, ])
130 FALSmq28_chr16.dat <- rbind(dat[1, ], dat[31872:33201, ], dat[39946, ],
        dat[39947, ])
131 FALSmq28_chr17.dat <- rbind(dat[1, ], dat[33202:34450, ], dat[39946, ],
        dat[39947, ])
132 FALSmq28_chr18.dat <- rbind(dat[1, ], dat[34451:35617, ], dat[39946, ],
        dat[39947, ])
133 FALSmq28_chr19.dat <- rbind(dat[1, ], dat[35618:36643, ], dat[39946, ],
        dat[39947, ])
134 FALSmq28_chr20.dat <- rbind(dat[1, ], dat[36644:37618, ], dat[39946, ],
        dat[39947, ])
135 FALSmq28_chr21.dat <- rbind(dat[1, ], dat[37619:38267, ], dat[39946, ],
        dat[39947, ])
136 FALSmq28_chr22.dat <- rbind(dat[1, ], dat[38268:38912, ], dat[39946, ],
        dat[39947, ])
137 FALSmq28_chr23.dat <- rbind(dat[1, ], dat[38913:39945, ], dat[39946, ],
        dat[39947, ])
138
139 # write these to file
140 write.table(FALSmq28_chr1.dat, file = "FALSmq28_chr1.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
141 write.table(FALSmq28_chr2.dat, file = "FALSmq28_chr2.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
142 write.table(FALSmq28_chr3.dat, file = "FALSmq28_chr3.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
143 write.table(FALSmq28_chr4.dat, file = "FALSmq28_chr4.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
144 write.table(FALSmq28_chr5.dat, file = "FALSmq28_chr5.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
145 write.table(FALSmq28_chr6.dat, file = "FALSmq28_chr6.dat.txt", sep = "\t",
        col.names = FALSE, row.names = FALSE)
```

```
write.table(FALSmq28_chr7.dat, file = "FALSmq28_chr7.dat.txt", sep = "\t",
    col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr8.dat, file = "FALSmq28_chr8.dat.txt", sep = "\t",
    col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr9.dat, file = "FALSmq28_chr9.dat.txt", sep = "\t",
    col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr10.dat, file = "FALSmq28_chr10.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr11.dat, file = "FALSmq28_chr11.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr12.dat, file = "FALSmq28_chr12.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr13.dat, file = "FALSmq28_chr13.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr14.dat, file = "FALSmq28_chr14.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr15.dat, file = "FALSmq28_chr15.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr16.dat, file = "FALSmq28_chr16.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr17.dat, file = "FALSmq28_chr17.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr18.dat, file = "FALSmq28_chr18.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr19.dat, file = "FALSmq28_chr19.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr20.dat, file = "FALSmq28_chr20.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr21.dat, file = "FALSmq28_chr21.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr22.dat, file = "FALSmq28_chr22.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
write.table(FALSmq28_chr23.dat, file = "FALSmq28_chr23.dat.txt", sep =
    "\t", col.names = FALSE, row.names = FALSE)
```

A.2.17 Running genome-wide linkage analyis using Merlin software

This script was used to run linkage analysis using Merlin software for family FALSmq28, for each chromosome.

```
#!/bin/bash

#  merlin_FALSmq28_chr1.sh

###
#script to run linkage analysis on FALSmq28 chr1
###

## set the working directory
cd /datastore/mcc549/FALSmq28_linkage

# load the merlin module
module load merlin/1.1.2

# change the destination of the temporary files generated by merlin
export TMPDIR=$JOBDIR

# run merlin
merlin -d ./chr1/FALSmq28_chr1.dat -p ./chr1/FALSmq28_chr1.ped -m
    ./chr1/FALSmq28_chr1.map --model parametric_new.model --bits 50 -fe
    --simwalk2 --swap --prefix chr1_merlin_out --pdf --tabulate
```

A.2.18 Analysis and plotting of results from genetic linkage analysis of FALSmq28

This script was used to analyse and plot the LOD score results from genetic linkage analysis of family FALSmq28 using Merlin.

```
# mq28_merlin_result_analysis.R

# this code is for analysing the results of merlin parametric linkage
    analysis for mq28

#laod required pacakges
library(dplyr)
library(ggplot2)

# first set the working directory
setwd("/Volumes/Personal/Bioinformatics/Linkage/QC and analysis")

# load the SNP chip annotation file from macrogen
annotation <- read.delim("/Volumes/Personal/Bioinformatics/Linkage/Raw\
    data/170123-InfiniumCoreExome-24_v1.1_Gene_annotation.txt")
View(annotation)
dim(annotation)
##[1] 551839     9

# load the merlin parametric linkage analysis results table for mq28
mq28.result <- read.delim("/Volumes/Personal/Bioinformatics/Linkage/Raw\
    data/merlin_results/mq28_merlin-parametric.tbl")
View(mq28.result)
dim(mq28.result)
##[1] 39944    7

# merge by marker name
mq28.result.w.annotation <- merge(mq28.result, annotation, by.x = "LABEL",
    by.y = "Name", all.x = TRUE)
dim(mq28.result.w.annotation)
##[1] 39944   15
View(mq28.result.w.annotation)

# sort by LOD score (descending)
```

```
mq28.result.w.annotation.ordered <-
    mq28.result.w.annotation[with(mq28.result.w.annotation, order(-LOD)), ]
View(mq28.result.w.annotation.ordered)
#subset the highest LOD scores
toptop.LOD <- subset(mq28.result.w.annotation.ordered,
    mq28.result.w.annotation.ordered$LOD > 1)
View(toptop.LOD)
colnames(toptop.LOD)
toptop.LOD$MODEL <- NULL
toptop.LOD$POS <- NULL
toptop.LOD$ALPHA <- NULL
toptop.LOD$HLOD <- NULL
toptop.LOD$Chr <- NULL
toptop.LOD$Genetic.Dist <- NULL
toptop.LOD$Mutation.s. <- NULL
View(toptop.LOD)
toptop.LOD <- toptop.LOD[c(1,2, 4, 3, 5:8)]
View(toptop.LOD)
write.csv(toptop.LOD, file = "top.LOD.csv")

# sort by Chr and Genetic.Dist
mq28.result.w.annotation$CHR <- as.character(mq28.result.w.annotation$CHR)
mq28.result.w.annotation$CHR <- as.numeric(mq28.result.w.annotation$CHR)
mq28.result.w.annotation.by.CHR <-
    mq28.result.w.annotation[with(mq28.result.w.annotation, order(CHR,
    Genetic.Dist)), ]
View(mq28.result.w.annotation.by.CHR)

# make a pretty graph

# make a new data frame to make our graph with
graph.data <- select(mq28.result.w.annotation.by.CHR, CHR, Genetic.Dist,
    MapInfo, LOD)
graph.data$chromosome <- graph.data$CHR
graph.data$chromosome <- as.character(graph.data$chromosome)
graph.data$chromosome <- as.factor(graph.data$chromosome)
View(graph.data)

# plot lod scores
```

```
ggplot(graph.data, aes(Genetic.Dist, colour = chromosome)) +
  geom_line(aes(y = LOD)) +
  facet_grid(~CHR, scales = "fixed", space = "free_x") +
  theme_bw() +
  theme(panel.spacing.x=unit(0, "lines")) +
  geom_hline(yintercept=0) +
  geom_hline(yintercept=1) +
  theme(legend.position="none") +
  scale_size_area(max_size = max(graph.data$LOD))
```

A.2.19 Functional distribution of WES and WGS variants from FALSmq28

This script was used to determine how the variants identified in FALSmq28 from WES or WGS data were distirbuted across the functional classes of the genome, and create stacked bar charts to reflect this distribution.

```
# mq28_func_class.R

# this script is to evaluate the number of mq28 variants falling into the different genomic functional categories from WES and WGS data

# load required packages
library(readr)
library(dplyr)
library(ggplot2)
library(data.table)

# set working directory
setwd("/Volumes/data_FMHS/Restrict/Blair Group/Emily/FALSmq28/Func.refGene")

# first import the tab delimited ANNOVAR annotated combined family VCF
mq28 <- read_delim("FALSmq28exomes_anno.hg19_multianno.txt", delim = "\t") # 185 703 lines

# tally classes
WES_Func.refGene_table <- as.data.frame(tally(group_by(mq28, Func.refGene)))

# calculate proportions
WES_Func.refGene_table$portion <- ((WES_Func.refGene_table$n)/(sum(WES_Func.refGene_table$n)))

# calculate percents
WES_Func.refGene_table$percent <- WES_Func.refGene_table$portion*100

# round off percents
WES_Func.refGene_table$percent_rounded <- round(WES_Func.refGene_table$percent, digits = 3)

# make stacked bar chart
```

```
x <- ggplot(WES_Func.refGene_table, aes(x = "", y = percent, fill =
    Func.refGene)) +
  geom_col(width = 1) + # width = 1 gets rid of the circle in the middle
  coord_polar("y") +
  theme_void()
ggsave(filename = "mq28_WES_Func.refGene_piechart.png", plot=x, device =
    "png", dpi = 600, units = "cm", height = 20, width = 20)

# repeat for ExonicFunc.refGene
# repeat for WGS
```

A.2.20 Identifying discordant variants between co-twins in WGS data

This Python script was written by Denis Bauer and edited by Natalie Twine, and was used to identify discordant variants between monozygotic twin pairs form a two-sample WGS VCF.

```
#discandfilter3.py

import sys,argparse

parser = argparse.ArgumentParser()
parser.add_argument('-i', dest='fileIN', help='The input fastq file')
parser.add_argument('-o', dest='fileOUT', help='The output VCF file')
parser.add_argument('-l', dest='locprivate', action='store_true',
    help='location cannot be discordant in other Twins')
parser.add_argument('-g', dest='gtprivate', action='store_true', help='One
    of the discordant GT must not be seen in the other Twins')
parser.add_argument('-c', dest='coverage', type=int, help='Coverage at
    which to accept the GT')
parser.add_argument('-m', dest='macrogen', action='store_true',
    help='Macrogen')

args = parser.parse_args()
print args
fileOUT=open(args.fileOUT,"w")
hashname={}

# Get properties of the GT entry for individual <count>
# returns [Genotype,Depth], e.g. [1/1,65]
def getProp(stringline,count):
    stringline=stringline.strip()
    arr=stringline.split("\t")
    if arr[count].split(":")[0]=="./." or arr[count].split(":")[0]==".":
        return[0,0]
    else:
        gt=arr[count].split(":")[0]
        dp=arr[count].split(":")[2]
        if dp==".":
            dp=0
```

```
30          return[gt,dp]
31
32  # Get the genotypes from the other Twin individuals in the file
33  # returns Set(Genotypes), e.g. (1/1,./.,0/0)
34  def genotypelist(stringline, count):
35      stringline=stringline.strip()
36      arr=stringline.split("\t")
37      othergt=[]
38      for i in range(9,len(arr)):
39          if i == count:
40              continue
41          othergt.append(arr[i].split(":")[0])
42      return set(othergt)
43
44  # datastructure indivating the Twin pairs and their location in the VCF
        filed
45  # it also has a counter variable that gets incremented if this pair had a
46  # discordant variant
47  #pairs=[["151002_FR07935773","151002_FR07935774",0,"set3"],
48  #["151002_FR07935768","151002_FR07935769",0,"set1"],
49  #["151002_FR07935770","151002_FR07935772",0,"set2"],
50  #["160215_FR07935864","160215_FR07935865",0,"set4"]]
51
52  #pairs=[["160215_FR07935864","151002_FR07935768",0,"set1"],
53  #["160215_FR07935865","151002_FR07935772",0,"set2"],
54  #["151002_FR07935769","151002_FR07935770",0,"set3"],
55  #["151002_FR07935773","151002_FR07935774",0,"set4"]]
56
57  #set 1 (triplet pair 1) - sets 1,2,3,4 that were run in previous analysis -
        set 5 new
58  #pairs=[["WIL1636A1","WIL907A2",0,"set1"],
59  #["WIL1636A2","WIL907A6",0,"set2"], #triplet pair 1
60  #["WIL907A3","WIL907A4",0,"set3"],
61  #["WIL908A1","WIL908A2",0,"set4"],
62  #["MQ160057","MQ160059",0,"set5"]]
63
64  #set 2 (triplet pair 2) - sets 1,3,4 same as previous analysis - set2 is
        alternate triplet pair - set 5 new
65  pairs=[["WIL1636A1","WIL907A2",0,"set1"],
```

```
["WIL1636A2","130-990370",0,"set2"], #triplet pair 2
["WIL907A3","WIL907A4",0,"set3"],
["WIL908A1","WIL908A2",0,"set4"],
["MQ160057","MQ160059",0,"set5"]]

if (args.macrogen):
    print("Macrogen")
    ############
    # Macrogen
    ############
    pairs=[["mqX-MQ150099","mqX-MQ150189",0,"set1"]]

# Main function
for i in open(args.fileIN):
    i=i.strip()
    if i[0:6]=="#CHROM":
        c=0
        for n in i.split("\t")[9::]:
            hashname[n.split(".")[0]]=c+9
            c+=1
        fileOUT.write("##INFO=<ID=discordant,Number=1,Type=String,
      Description=\"Discordant setID\">\n")
        fileOUT.write(i+"\n")
        continue

    if i[0]=="#":
        fileOUT.write(i+"\n")
        continue

    discordant=[]

    # iterate through Twins
    for p in range(0,len(pairs)):

        # Get genotype and depth
        p1=getProp(i,hashname[pairs[p][0]])
        p2=getProp(i,hashname[pairs[p][1]])

```

```
104         # Find out if it is discordant: 1) GT were called for both 2) GT
                differ 3) Coverage above <coverage>
105         if ((p1[0]!=0 and p2[0]!=0) and (p1[0]!=p2[0]) and
                (int(p1[1])>args.coverage and int(p2[1])>args.coverage)):
106 #           print "%i %i %s" % (hashname[pairs[p][0]],
    hashname[pairs[p][1]], i)
107             if (not(args.gtprivate) or (not(p1[0] in
                    genotypelist(i,hashname[pairs[p][0]])) or not(p2[0] in
                    genotypelist(i,hashname[pairs[p][1]])))):
108                 discordant.append(p)
109 #           if (((p1[0] in genotypelist(i,hashname[pairs[p][0]])) or (p2[0]
    in genotypelist(i,hashname[pairs[p][1]])))):
110 #               print "non priv"
111 #               die
112
113     # if this line had at least one discordant pair print it
114     if len(discordant)>0:
115
116         # if they need to be private to the Twin pair do not print
117         if args.locprivate and len(discordant)>1:
118             continue
119
120         d=[]
121         for x in discordant:
122             d.append(pairs[x][3]+"_"+pairs[x][0]+"_"+pairs[x][1]) #prep
                    string
123             pairs[x][2]+=1 # count
124
125         arr=i.split("\t")
126         arr[7]+=";discordant="+",".join(d)
127         fileOUT.write("\t".join(arr)+"\n")
128
129 for i in pairs:
130     print " | ".join(map(str,i))
131
132 # Close files
133 fileOUT.close()
```

A.2.21 Identifying WGS discordant variants also genotyped by a SNP microarray

This script was used to determine which variants that had been identified as discordant between monozygotic co-twins using WGS data, had also been genotyped using the Illumina Infinium CoreExome-24 BeadChip v1.0 or v1.1 microarray.

```
#!/bin/sh -e

#  twin_disc_variant_validation_SNPchip.sh

# This code is for determine whether any discordant variants identified
    between co-twins/triplets had also been genotyped by the microarray, and
    to subsequently extract and compare the associated genotypes.

# set working directory
cd /Volumes/twin_discordant_variants

# bgzip the VCFs so we can merge them with bcftools
bgzip -c ./set1/WIL1636A1_WIL907A2.vqsr.vep.vcf >
    ./set1/WIL1636A1_WIL907A2.vqsr.vep.vcf.gz
bgzip -c ./set1/WIL1636A1_WIL907A2L.vqsr.vep.vcf >
    ./set1/WIL1636A1_WIL907A2L.vqsr.vep.vcf.gz
bgzip -c ./set1/WIL1636A1_WIL907A2LG.vqsr.vep.vcf >
    ./set1/WIL1636A1_WIL907A2LG.vqsr.vep.vcf.gz

bgzip -c ./set2/WIL1636A2_WIL907A6.vqsr.vep.vcf >
    ./set2/WIL1636A2_WIL907A6.vqsr.vep.vcf.gz
bgzip -c ./set2/WIL1636A2_WIL907A6L.vqsr.vep.vcf >
    ./set2/WIL1636A2_WIL907A6L.vqsr.vep.vcf.gz
bgzip -c ./set2/WIL1636A2_WIL907A6LG.vqsr.vep.vcf >
    ./set2/WIL1636A2_WIL907A6LG.vqsr.vep.vcf.gz

bgzip -c ./set3/WIL907A3_WIL907A4.vqsr.vep.vcf >
    ./set3/WIL907A3_WIL907A4.vqsr.vep.vcf.gz
bgzip -c ./set3/WIL907A3_WIL907A4L.vqsr.vep.vcf >
    ./set3/WIL907A3_WIL907A4L.vqsr.vep.vcf.gz
bgzip -c ./set3/WIL907A3_WIL907A4LG.vqsr.vep.vcf >
    ./set3/WIL907A3_WIL907A4LG.vqsr.vep.vcf.gz

```

```
bgzip -c ./set4/WIL908A1_WIL908A2.vqsr.vep.vcf > ./set4/WIL908A1_WIL908A2.vqsr.vep.vcf.gz
bgzip -c ./set4/WIL908A1_WIL908A2L.vqsr.vep.vcf > ./set4/WIL908A1_WIL908A2L.vqsr.vep.vcf.gz
bgzip -c ./set4/WIL908A1_WIL908A2LG.vqsr.vep.vcf > ./set4/WIL908A1_WIL908A2LG.vqsr.vep.vcf.gz

bgzip -c ./set5/MQ160057_MQ160059.vqsr.vep.vcf > ./set5/MQ160057_MQ160059.vqsr.vep.vcf.gz
bgzip -c ./set5/MQ160057_MQ160059L.vqsr.vep.vcf > ./set5/MQ160057_MQ160059L.vqsr.vep.vcf.gz
bgzip -c ./set5/MQ160057_MQ160059LG.vqsr.vep.vcf > ./set5/MQ160057_MQ160059LG.vqsr.vep.vcf.gz

# index these VCFs so we can merge them with bcftools
tabix -p vcf ./set1/WIL1636A1_WIL907A2.vqsr.vep.vcf.gz
tabix -p vcf ./set1/WIL1636A1_WIL907A2L.vqsr.vep.vcf.gz
tabix -p vcf ./set1/WIL1636A1_WIL907A2LG.vqsr.vep.vcf.gz

tabix -p vcf ./set2/WIL1636A2_WIL907A6.vqsr.vep.vcf.gz
tabix -p vcf ./set2/WIL1636A2_WIL907A6L.vqsr.vep.vcf.gz
tabix -p vcf ./set2/WIL1636A2_WIL907A6LG.vqsr.vep.vcf.gz

tabix -p vcf ./set3/WIL907A3_WIL907A4.vqsr.vep.vcf.gz
tabix -p vcf ./set3/WIL907A3_WIL907A4L.vqsr.vep.vcf.gz
tabix -p vcf ./set3/WIL907A3_WIL907A4LG.vqsr.vep.vcf.gz

tabix -p vcf ./set4/WIL908A1_WIL908A2.vqsr.vep.vcf.gz
tabix -p vcf ./set4/WIL908A1_WIL908A2L.vqsr.vep.vcf.gz
tabix -p vcf ./set4/WIL908A1_WIL908A2LG.vqsr.vep.vcf.gz

tabix -p vcf ./set5/MQ160057_MQ160059.vqsr.vep.vcf.gz
tabix -p vcf ./set5/MQ160057_MQ160059L.vqsr.vep.vcf.gz
tabix -p vcf ./set5/MQ160057_MQ160059LG.vqsr.vep.vcf.gz


# merge discordant variant VCFs from each twin/triplet pair to create a combined VCF for all discordant variants across twin sets
```

```
bcftools merge ./set1/WIL1636A1_WIL907A2.vqsr.vep.vcf.gz
    ./set2/WIL1636A2_WIL907A6.vqsr.vep.vcf.gz
    ./set3/WIL907A3_WIL907A4.vqsr.vep.vcf.gz
    ./set4/WIL908A1_WIL908A2.vqsr.vep.vcf.gz
    ./set5/MQ160057_MQ160059.vqsr.vep.vcf.gz >
    AllTwinSets_discordant.vqsr.vep.vcf.gz

#subset rsID variants
bcftools view -Ov -i'ID!="."' AllTwinSets_discordant_variants.vqsr.vep.vcf
    > AllTwinSets_discordant_variants_rsID_only.vqsr.vep.vcf
```

A.2.22 Extracting SNP microarray genotypes for WGS-derived discordant variants

This script was written by Kelly Williams and was used to extract the SNP microarray genotype data for all tiwns/triplets, for the 81 putative discordant variants identified by WGS data that had been genotyped using the Illumina Infinium CoreExome-24 BeadChip v1.0 or v1.1 microarray.

```
# twin_disc_variant_validation_SNPchip.R

# identifying discordant SNPs in existing Illumina Infinium CoreExome-24
    chips
# written by Kelly Williams

discordantRSID <- read.csv("~/Downloads/discordantRSID.csv")
SNPfile <- read.delim("~/Downloads/SNPfile.txt")
InfiniumCoreExome.24_v1.1_Gene_annotation <-
    read.delim("~/Downloads/170123-InfiniumCoreExome-24_v1.1_Gene_annotation.
    txt")
AllTwinSets_discordant_analysis1_analysis2.vqsr.vep <-
    read.table("~/Downloads/AllTwinSets_discordant_analysis1_analysis2.vqsr.
    vep.vcf", quote="\"")
HumanCoreExome.24v1.0_A_annotated <-
    read.delim("~/Downloads/HumanCoreExome-24v1-0_A_annotated.txt")


which(SNPfile$Name %in% discordantRSID$ID)
which(InfiniumCoreExome.24_v1.1_Gene_annotation$Name %in% discordantRSID$ID)
```

```
length(which(SNPfile$Name %in% discordantRSID$ID))

x <-
    InfiniumCoreExome.24_v1.1_Gene_annotation[which(InfiniumCoreExome.24_v1.1
    _Gene_annotation$Name %in% discordantRSID$ID),]
y <- SNPfile[which(SNPfile$Name %in% discordantRSID$ID),]

SNPfile$location <- paste(SNPfile$Chr,":",SNPfile$Position, sep = "")
InfiniumCoreExome.24_v1.1_Gene_annotation$location <-
    paste(InfiniumCoreExome.24_v1.1_Gene_annotation$Chr,":",
    InfiniumCoreExome.24_v1.1_Gene_annotation$MapInfo, sep = "")
discordantRSID$location <-
    paste(discordantRSID$CHROM,":",discordantRSID$POS, sep = "")
AllTwinSets_discordant_analysis1_analysis2.vqsr.vep$location <-
    paste(AllTwinSets_discordant_analysis1_analysis2.vqsr.vep$V1,":",
    AllTwinSets_discordant_analysis1_analysis2.vqsr.vep$V2, sep = "")

length(which(InfiniumCoreExome.24_v1.1_Gene_annotation$location %in%
    discordantRSID$location))

z <- InfiniumCoreExome.24_v1.1_Gene_annotation[which(InfiniumCoreExome.
    24_v1.1_Gene_annotation$location %in% discordantRSID$location),]

discordant_positions <-
    as.vector(AllTwinSets_discordant_analysis1_analysis2.vqsr.vep$location)

length(which(InfiniumCoreExome.24_v1.1_Gene_annotation$location %in%
    discordant_positions))

w <- InfiniumCoreExome.24_v1.1_Gene_annotation[which(InfiniumCoreExome.
    24_v1.1_Gene_annotation$location %in% discordant_positions),]
v <- SNPfile[which(SNPfile$location %in% discordant_positions),]

#doublechecking
which(AllTwinSets_discordant_analysis1_analysis2.vqsr.vep$V3 ==
    "variant.67526")

#write files
setwd("~/Desktop")
```

```
write.csv(v, "Discordant_SNPs_on_CoreExome.csv", quote = FALSE, row.names = FALSE)
write.csv(w, "Discordant_SNPs_on_CoreExome_annotated.csv", quote = FALSE, row.names = FALSE)

#import full data to subset out twins
samples <- read.csv("~/Desktop/MWAS_GWAS_WGS_master_sample_sheet_all_data-IT_5-2-18.csv")
twins <- samples[-which(samples$Twin.code.deidentified == ""),]

#remove PSP twins and control twins that were not WGS
twins <- twins[-which(twins$MQ.UniqueID == "gleher11945"),]
twins <- twins[-which(twins$MQ.UniqueID == "glewal11945"),]
twins <- twins[-which(twins$MQ.UniqueID == "gilfle21969"),]
twins <- twins[-which(twins$MQ.UniqueID == "gorsuz21969"),]

#generate list for Ruqian to extract the twins from the rest of the data and generate a genotype file
twin_Ruqian <- twins[,c(1,4:6,67:74)]
twin_Ruqian <- twin_Ruqian[-which(twin_Ruqian$HumanCoreExome.24v1.0_A == ""),]
write.csv(twin_Ruqian, "twin_list_for_genotypes.csv", quote = FALSE, row.names = FALSE)

#Ruqian has pulled out all genotypes for the twins using Illumina's Genome Studio
# Import and then subset based on the discordant SNPs
alltwinMatched1 <- read.csv("~/Downloads/twinsGtype/alltwinsSignal/alltwinMatched1.txt", sep="")
alltwinMatched2 <- read.delim("~/Downloads/twinsGtype/alltwinsSignal/alltwinMatched2.txt")
alltwinMatched1_2 <- read.table("~/Downloads/twinsGtype/alltwinsSignal/alltwin1.txt", quote="\"", comment.char="")

alltwin1 <- merge(x = alltwinMatched1_2, y = alltwinMatched1, by.x = "V2", by.y = "SampleID")
```

```
alltwin1 <- merge(x = alltwin1, y = twins[,c(4:6)], by.x = "V1", by.y =
    "Sample.ID", all.x = TRUE)
alltwin1$inputID <- paste("sample.",alltwin1$V2,sep = "")

#use terminal
#paste sample.1000377 sample.1000378 sample.1000380 sample.1000381
    sample.1000382 sample.1000383 sample.1000384 sample.1000385
    sample.1000386 sample.1000387 sample.1000391 sample.1000392
    sample.1000397 sample.1000398 sample.1000400 sample.1000401
    sample.1000402 sample.1000403 sample.1000405 sample.1000406
    sample.1000407 sample.1000408 sample.1000409 sample.1000410
    sample.1000411 sample.1000412 sample.1000413 sample.1000414 sample.130
    sample.130_R1 sample.130_R2 sample.6 sample.6_R1 sample.6_R3 >
    combined.txt

all.genotypes <-
    read.delim("~/Downloads/twinsGtype/alltwinsSignal/combined.txt")
all.genotypes$location <-
    paste(all.genotypes$Chr,":",all.genotypes$Position, sep = "")
genotypes.subset <- all.genotypes[which(all.genotypes$location %in%
    discordant_positions),]

#change any not called (NC) to NA
genotypes.subset[ genotypes.subset == "NC" ] <- NA

row.names(genotypes.subset) <- genotypes.subset$Name

#get rid of duplicate columns
genotypes.subset <- genotypes.subset[, -grep("Name",
    colnames(genotypes.subset))]
genotypes.subset <- genotypes.subset[, -grep("Allele.",
    colnames(genotypes.subset))]
genotypes.subset <- genotypes.subset[, -grep("Chr.",
    colnames(genotypes.subset))]
genotypes.subset <- genotypes.subset[, -grep("Position.",
    colnames(genotypes.subset))]
genotypes.subset <- genotypes.subset[, -grep("Log.R.",
    colnames(genotypes.subset))]

```

```
genotypes.subset <- as.data.frame(t(genotypes.subset[,-c(1:2,37)]))
#genotypes.subset <- as.data.frame(genotypes.subset[,-c(1:2,37)])

genotypes.subset$sample <- rownames(genotypes.subset)
genotypes.subset$sample <- gsub("X","",genotypes.subset$sample)
genotypes.subset$sample <- gsub(".GType","",genotypes.subset$sample)

a <- merge(x = genotypes.subset, y = alltwin1, by.x = "sample", by.y =
    "V2", all.x = TRUE)

# inspect data manually
```

A.2.23 Determining the distribution of discordant variants bewteen SNP and indel variant types

This script was used to first annotate each discordant variant (indentified from each twin set, and from each processed dataset) as a SNP or indel using SNPSift, and subsequently count the number of lines containing each annotation, to determine the distribution of discordant variants across these variant types.

```
#!/bin/sh

#  variant_type_count.sh

# must run in SnpEff folder
# run for all disc twin VCFs

# annotate with variant type using SNPSift VarType
java -jar SnpSift.jar varType
    /full_path/set1_discordant_variants.vqsr.vep.vcf.vartype.vcf >
    /full_path/set1_discordant_variants_VARTYPEanno.vqsr.vep.vcf.vartype.vcf

# extract variant type annotation column to separate text file
java -jar SnpSift.jar extractFields
    /full_path/set1_discordant_variants_VARTYPEanno.vqsr.vep.vcf.vartype.vcf
    VARTYPE > /full_path/set1_VARTYPE.txt

# count number of lines containing the given variant type ie. SNP and INS
    or DEL
grep -c SNP
    /full_path/set1_discordant_variants_VARTYPEanno.vqsr.vep.vcf.vartype.vcf
grep -c "INS\|DEL"
    /full_path/set1_discordant_variants_VARTYPEanno.vqsr.vep.vcf.vartype.vcf
```

A.2.24 Creating Venn diagrams of discordant variants from the four bioinformatics processing pipelines

This R script was used to generate Venn diagrams for the discordant variants identified by each of the four processing pipelines, for each twin pair.

```
# venn_diagrams.R

# This script is for creating Venn diagrams for the discordant variants
    identified between ALS discordant MZ co-twins from WGS using different
    bioinformatics processing

library(VennDiagram)

#set working directory
setwd("/full_path/Four_processed_discordant_variants/twins_WGS_R_project")

###################
### twin set 1 ###
###################
# Reference four-set diagram
venn.plot <- draw.quad.venn(
  area1 = 18599,
  area2 = 37226,
  area3 = 1976,
  area4 = 1833,
  n12 = 0,
  n13 = 0,
  n14 = 0,
  n23 = 124,
  n24 = 251,
  n34 = 38,
  n123 = 0,
  n124 = 0,
  n134 = 0,
  n234 = 8,
  n1234 = 0,
  category = c("A", "B", "C", "D"),
  fill = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF"),
  lty = "dashed",
```

```
  cex = 2,
  cat.cex = 2,
  cat.col = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF")
)
grid.draw(venn.plot);
grid.newpage();
png(filename = "set1_venn.png", width = 10, height = 10, units = "cm", res
    = 600)
grid.draw(venn.plot)
dev.off()

####################
### twin set 2A ###
####################
# Reference four-set diagram
venn.plot <- draw.quad.venn(
  area1 = 12240,
  area2 = 33430,
  area3 = 1947,
  area4 = 635,
  n12 = 0,
  n13 = 0,
  n14 = 0,
  n23 = 150,
  n24 = 107,
  n34 = 38,
  n123 = 0,
  n124 = 0,
  n134 = 0,
  n234 = 12,
  n1234 = 0,
  category = c("A", "B", "C", "D"),
  fill = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF"),
  lty = "dashed",
  cex = 2,
  cat.cex = 2,
  cat.col = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF")
)
grid.draw(venn.plot);
```

```
grid.newpage();
png(filename = "set2A_venn.png", width = 10, height = 10, units = "cm", res
    = 600)
grid.draw(venn.plot)
dev.off()

####################
### twin set 2B ###
####################
# Reference four-set diagram
venn.plot <- draw.quad.venn(
  area1 = 14097,
  area2 = 15577,
  area3 = 2010,
  area4 = 1088,
  n12 = 0,
  n13 = 0,
  n14 = 0,
  n23 = 66,
  n24 = 84,
  n34 = 49,
  n123 = 0,
  n124 = 0,
  n134 = 0,
  n234 = 4,
  n1234 = 0,
  category = c("A", "B", "C", "D"),
  fill = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF"),
  lty = "dashed",
  cex = 2,
  cat.cex = 2,
  cat.col = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF")
)
grid.draw(venn.plot);
grid.newpage();
png(filename = "set2B_venn.png", width = 10, height = 10, units = "cm", res
    = 600)
grid.draw(venn.plot)
dev.off()
```

```

##################
### twin set 4 ###
##################
# Reference four-set diagram
venn.plot <- draw.quad.venn(
  area1 = 55132,
  area2 = 157012,
  area3 = 6358,
  area4 = 7441,
  n12 = 0,
  n13 = 0,
  n14 = 0,
  n23 = 954,
  n24 = 2371,
  n34 = 458,
  n123 = 0,
  n124 = 0,
  n134 = 0,
  n234 = 263,
  n1234 = 0,
  category = c("A", "B", "C", "D"),
  fill = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF"),
  lty = "dashed",
  cex = 2,
  cat.cex = 2,
  cat.col = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF")
)
grid.draw(venn.plot);
grid.newpage();
png(filename = "set4_venn.png", width = 10, height = 10, units = "cm", res
    = 600)
grid.draw(venn.plot)
dev.off()

##################
### twin set 5 ###
##################
# Reference four-set diagram
```

```
venn.plot <- draw.quad.venn(
  area1 = 30994,
  area2 = 22755,
  area3 = 2646,
  area4 = 2480,
  n12 = 0,
  n13 = 0,
  n14 = 0,
  n23 = 110,
  n24 = 268,
  n34 = 75,
  n123 = 0,
  n124 = 0,
  n134 = 0,
  n234 = 12,
  n1234 = 0,
  category = c("A", "B", "C", "D"),
  fill = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF"),
  lty = "dashed",
  cex = 2,
  cat.cex = 2,
  cat.col = c("#009292", "#FFB6DB", "#B66DFF", "#6DB6FF")
)
grid.draw(venn.plot);
grid.newpage();
png(filename = "set5_venn.png", width = 10, height = 10, units = "cm", res
    = 600)
grid.draw(venn.plot)
dev.off()
```

A.3 Additional tables

A.3.1 Primer details

The following table contains details of each primer set designed and utilised as part of this thesis. Deatils include primer sequences, product sizes, optimised PCR conditions, type of sequencing performed on PCR products and the purpose of each primer set.

Table A.1: Primer details

Primer Name	Sequence	PCR product size	Optimised PCR conditions*	Ta	Sequencing	Purpose
SOD1_Ex1_NewF	ATTGGTTTGGGGCCAGAG	408	Touchdown PCR thermocycling	64	Sanger	ALS gene
SOD1_Ex1_NewR	TGACTCAGCACTTGGGCAC					
SOD1_Ex2_NewF	GTCAGCCTGGGATTTGGAC	355	Touchdown PCR thermocycling;	64	Sanger	ALS gene
SOD1_Ex2_NewR	CGACAGAGCAAGACCCTTTC		and PCR enhancer			
SOD1_Ex3_NewF	CAGAAGTCGTGATGCAGGTC	313	Standard	61	Sanger	ALS gene
SOD1_Ex3_NewR	CAGCAAGTTCAAAAGCAAAGG					
SOD1_Ex4_NewF	GACGTGAAGCCTTGTTTGAAG	418	Touchdown PCR thermocycling	59	Sanger	ALS gene
SOD1_Ex4_NewR	AATTGTCCAATAAAATTGCTTTT					
SOD1_Ex5_NewF	TTCATTTAGACAGCAACACTTACC	572	Standard	60	Sanger	ALS gene
SOD1_Ex5_NewR	CAAAATACAGGTCATTGAAACAGAC					
C9orf72_FAM_F	6-FAM-AGTCGCTAGAGGCGAAAGC	~300	50ng DNA; 0.2mM deazoGTP;	70	Fragment	ALS gene
C9orf72_A	TACGCATCCCAGTTTGAGACG		1M Betaine; DMSO;		length	
C9orf72_R	TACGCATCCCAGTTTGAGACGGGGGG		Touchdown PCR thermocycling		analysis	
	CCGGGGCCGGGGCCGGGG					
EEF1D_Ex8_F	gagcgagtcccagagtgaac	299	Standard	65	Sanger	Proband candidate
EEF1D_Ex8_R	aggatgactgtgtgggaacag					
SPTBN4_Ex30_F	CAGGGGAACAGCCATTG	197	Standard	65	Sanger	Proband candidate
SPTBN4_Ex30_R	TATAGAGCCATGGGTGTGGG					
ABCC2_Ex21_F	AGTGACTGTGACATCTGCTTGC	303	Standard	62	Sanger	Proband candidate
ABCC2_Ex21_R	TGTAAGTATGCGTTCAATTTTCAC					
ABCC2_Ex25_F	AAAGGAGGAAGATGGTGGATG	336	Standard	64	Sanger	Proband candidate
ABCC2_Ex25_R	CCCACCGCTAATATCAAACATATAG					
MTHFR_Ex10_F	ACCTTAGGTGTCTGCGAAAGG	238	Standard	61	Sanger	Proband candidate
MTHFR_Ex10_R	gctaggtgctgggtgtttg					
DAGLB_Ex12_13_M13_F	CTTTCATGGAAGCCCTTGTG	359	Standard	64	Sanger	Proband candidate
DAGLB_Ex12_13_M13_R	CCTCTCCACAGGATCTCAGGG					
TIA1_Ex3_F	TCCATTCATTCTCCATGGCCTGA	539	Standard	63	Sanger	Proband candidate
TIA1_Ex3_R	GAGTTTGAGACCAGTCTGGCT					
TIA1_Ex11_F	TTGTTTTGGCTAAGAATTTGTGG	345	Standard	63	Sanger	Proband candidate
TIA1_Ex11_R	TTACGCTTTACATAAGAGGCCC					
TIA1_Ex10_F	CAAGTTGCCCCAGAACTACAAG	451	Standard	63	Sanger	Proband candidate
TIA1_Ex10_R	CAATCCATGAAACACCATTCTG					
CLCN4_Ex12_F	GGGATTCTAGATGGTGTGTGTG	392	Standard	64	Sanger	FALS15
CLCN4_Ex12_R	CCTCCACAttcttcagggc					
SCN4A_Ex5_F	TCTGTCCTACCACCCACCC	232	Standard & PCR enhancer	63	Sanger	FALS15
SCN4A_Ex5_R	ACACTGAGTCAGGTTCCAGGC					
MTSS1L_Ex5_F	GCAGTTCACCAAgtgagtgg	284	Standard & PCR enhancer	66	Sanger	FALS15
MTSS1L_Ex5_R	TGTACTctgcagaaggggagag					
SUPV3L1_Ex4_F	agctactgtgcccagAGAGGAC	321	Standard	64	Sanger	FALS15
SUPV3L1_Ex4_R	TAATGACCACGAATCATCCAAG					
LRRN2_p.I196T_F	TCTCAACCACAACCAGCTCTAC	178	Standard	58	Sanger	FALS15
LRRN2_p.I196T_R	GTCCAGGATGGCATCTACCTT					
SPEG_p.P674R_F	GTACCCCAGACCTTGGAGAAG	160	Standard	64	Sanger	FALS15
SPEG_p.P674R_R	tacCGAGCTCAGGGGAGGT					
FAM171A1_Ex8_F	TGCTCTCACAGCCTTTATTTGA	184	Standard	64	Sanger	FALS15
FAM171A1_Ex8_R	GACGTAGGTCTCTCGAGGTGAT					
HOXD3_Ex3_F	TGGTGGAATTGGAAAAGGAAT	152	Standard	58	Sanger	FALS15

HOXD3_Ex3_R	CTTGGCCTTCTGGTCCTTCT					
MAPKAPK3_Ex11_F	TGGTTCCTAAGGTCAGTACATCC	276	Standard	64	Sanger	FALS15
MAPKAPK3_Ex11_R	GGCCTGAGCACATTTCAGTC					
SIM1_Ex11_F	ACCACCCTACTGTCTCTCCAAA	209	Standard	64	Sanger	FALS15
SIM1_Ex11_R	AGATGTTCCCTTGTGTCCTTGT					
TYMP_Ex6_F	AACTCTCCCAAGAAGCTCCAG	243	Standard	64	Sanger	FALS15
TYMP_Ex6_R	AGGGGTGAAGGGTAGGCTG					
SP1_p.A145T_F	GCTACCCCTACCTCAAAGGAAC	163	Standard	64	Sanger	FALS15
SP1_p.A145T_R	TAGGCATCACTCCAGGTAGTCC					
ZNF385B_Ex10_F	TTATGAAAAGATGCCTGTGGTG	359	Standard	62	Sanger	FALS15
ZNF385B_Ex10_R	TGAATCCTTGTGGCTTTCTTTC					
NECAB3_Ex7_F	agtcctcggccctgtgag	244	Standard	63	Sanger	FALS15
NECAB3_Ex7_R	atgaagtgaggggcagtgaag					
CEP295_p.N1707I_F	GTGATCCCAGGGTTTCAAGATA	185	Standard	64	Sanger	FALS15
CEP295_p.N1707I_R	TGCAATGCTGTTTGTCTCTGTA					
TNS2_p.S992L_F	CTAGCCCAGTCTCTCCGACCT	189	Standard	64	Sanger	FALS15
TNS2_p.S992L_R	GAGTGAGGGGAGACCCATCT					
ZNF425_p.R424P_F	AGATTAAGCTGGACGAGCACAT	179	Standard	61	Sanger	FALS15
ZNF425_p.R424P_R	CATGGCGTTCCTCCAGAAG					
ZNF497_Ex3_pt1_F	ggtacttgccttttctctcctg	183	Standard	64	Sanger	FALS15
ZNF497_Ex3_pt1_R	AACCTCCGTGGAGTTTTCC					
RNF133_p.R94Q_F	GGAGTTATAGTGCCACCAGAGG	232	Standard	58	Sanger	FALS15
RNF133_p.R94Q_R	CATCTTCAAATGCCTGATGAAA					
GDPD1_Ex7_F	TGTCTTGGGAAATACTGAGAGTTG	335	Standard	64	Sanger	FALS45
GDPD1_Ex7_R	TGGAAATCACTACAGAAATCTCTTC					
GPX7_Ex1_F	CCTGCGGAGGGAACGAG	380	Standard & PCR enhancer	66	Sanger	FALS45
GPX7_Ex1_R	GTCCTCGGAGCCACACC					
SCCPDH_Ex7_F	GCTTATTCAGACATTTACCAGCC	256	Standard	64	Sanger	FALS45
SCCPDH_Ex7_R	CCCACACTGAATAGAAAGAGGAAC					
PVRL3_Ex2_F	gatagttacacaggggtcagg	473	Standard	64	Sanger	FALS45
PVRL3_Ex2_R	tcttcaccactatcaccaaaataca					
GABRG3_Ex6_F	GGCCTAAAAGTTTAACTCCTAACTCC	362	Standard	64	Sanger	FALS45
GABRG3_Ex6_R	CACTTATGTATCATGGTTGCCC					
ARAP3_p.G330A_F	tcttataaaatctggggcagga	168	Standard	64	Sanger	FALS45
ARAP3_p.G330A_R	TGATGACCTGGAACTTGTTGTC					
KRT85_p.S5P_F	gctttccactcctttatgcag	219	Standard	64	Sanger	FALS45
KRT85_p.S5P_R	AGCTGAGCAGGAGCTGAAGTT					
DMWD_p.G425S_F	CTTTGACCCCTACACCACAAG	228	Standard	64	Sanger	FALS45
DMWD_p.G425S_R	TAGAGCACGTCTTCAGTGAGGT					
LZTFL1_Ex2_F	atctagcagtctcgaccacaGG	315	Standard	64	Sanger	FALS45
LZTFL1_Ex2_R	ATTGTCTGGCCTCTGCTATGG					
SLC22A5_Ex8_F	TTTGTTTTGCTCTCAATAGCTG	313	Standard	62	Sanger	FALS45
SLC22A5_Ex8_R	AAGCCAGTTAGTACTTCCATCCC					
GRIN2D_p.V144L_F	CCCATCCTCGACTTCCTGT	245	Standard & PCR enhancer	65	Sanger	FALS45
GRIN2D_p.V144L_R	ctaagccctgcctaccatctg					
SPATA2_p.G206S_F	CAAGTGAAGGACAAGGGCTACT	239	Standard	64	Sanger	FALS45
SPATA2_p.G206S_R	CAGTAGCTGTCATAGGCATCCA					
HIST1H3G_p.P39S_F	AGACTGCACGCAAGTCCAC	168	Standard	64	Sanger	FALS45
HIST1H3G_p.P39S_R	GCAGCTCAGTCGACTTCTGATA					
PIGZ_Ex2_F	TGACAGATCCATTTTCAGTTTG	391	Standard	64	Sanger	FALS45
PIGZ_Ex2_R	TACCCAGACATGCTACTCCCTC					
NPBWR1_p.L252V_F	GTCCTCTATACCACCCTGCTGT	224	Standard	65	Sanger	FALS45

NPBWR1_p.L252V_R	AGGCTGGTGATGAAGTAGGAGA					
ASXL2_p.A796T_F	CTGGAGCACAACTACAGCAAAC	197	Standard	64	Sanger	FALS45
ASXL2_p.A796T_R	AGAAGGTGCTTTCTCCTGTCTG					
ORM1_p.K138N_F	CTAGGCCTCCTCACCTGTAAGA	195	Standard	64	Sanger	FALS45
ORM1_p.K138N_R	gcatgcctacCATAGACAGACA					
ZNF132_p.G455R_F	AACAATAACTCCAACCTTGCTCA	225	Standard	64	Sanger	FALS45
ZNF132_p.G455R_R	ATAAGGCCTTTGCCCAGTATGT					
CROCC_p.A382S_F	CCTACGGCTGGCAGAGAG	227	Standard & PCR enhancer	60	Sanger	FALS45
CROCC_p.A382S_R	TCTCCCCAACTCTCCAGTTTTA					
FHAD1_Ex15_F	GGTAAACAAATGGGGAAAGACTC	298	Standard	64	Sanger	FALS45
FHAD1_Ex15_R	CTCAATTTCCCATAGAAAAGGC					
STRN4_Ex8_F	ACCTATGCCAGACTGGGTTG	240	Standard & PCR enhancer	54	Sanger	FALSmq2
STRN4_Ex8_R	CAGCTCTGCAGCCTCCC					
LZTR1_Ex14_F	GTGAGGTGCCTAACCGCC	343	Standard & PCR enhancer	66	Sanger	FALSmq2
LZTR1_Ex14_R	atcagtaaggcagggctgg					
EHBP1_p.Q654L_F	TTGGAGAGTCAGAAAGTGAGCA	179	Standard	64	Sanger	FALSmq2
EHBP1_p.Q654L_R	TGCTTGGGTTGAATCTGTATTG					
EMP2_Ex5_F	GTCTTCAACTCTGGCCGTATG	511	Standard	60	Sanger	FALSmq2
EMP2_Ex5_R	GCAGTTCTGAATACCAGCCTTC					
EXOC3L1_Ex7_F	tcagatccctgctacattcctt	243	Standard	64	Sanger	FALSmq2
EXOC3L1_Ex7_R	aaagcctccctccttcctct					
TUSC5_Ex2_F	AAGCTGACCCACAGCCTTC	255	Standard	62	Sanger	FALSmq2
TUSC5_Ex2_R	GGTACACCCTTGAGCAGTCC					
DPH6_Ex7_F	gctgtctaattttaaacttctttcttg	210	Standard	57	Sanger	FALSmq2
DPH6_Ex7_R	tgttcagttgcttccatcactt					
CHRNA2_p.E411Q_F	CCTCTTATCACTGGCTGGAGAG	192	Standard & PCR enhancer	63	Sanger	FALSmq2
CHRNA2_p.E411Q_R	ATAGCAGCAGCTCACCCTCCT					
FAM205A_p.K1282N_F	GGGGAGAAGTAAGACGGAGAAG	140	Standard	64	Sanger	FALSmq2
FAM205A_p.K1282N_R	AGCCAGGTTGTCTGGAGTGTAG					
P2RY2_p.W16R_F	cattctcaaggttccagagctt	250	Standard	64	Sanger	FALSmq2
P2RY2_p.W16R_R	CACGTACTTGAAGTCCTCGTTG					
ALPK1_p.D979N_F	CTGAATTCCAGTGGGAGTTCTT	187	Standard	64	Sanger	FALSmq2
ALPK1_p.D979N_R	tacTATGTGCTCGGTGGAGTTG					
SLC25A21_Ex7_F	GCTTGAAGGGAAGTTAATACGG	345	Standard	64	Sanger	FALSmq2
SLC25A21_Ex7_R	cagcacagtgaccaggacag					
ZFHX2_p.T565S_F	GCTGTGACGTCTGCAACTACTC	225	Standard & PCR enhancer	61	Sanger	FALSmq2
ZFHX2_p.T565S_R	GGTTACGGGAGATGTTTGTCTC					
CFH_p.A421G_F	tcattgttatggtccttagGAAA	189	Standard	58	Sanger	FALSmq2
CFH_p.A421G_R	GATGCATCTGGGAGTAGGAGAC					
PCDHB11_p.S759T_F	CTTTCCCAGAGCTACCAGTACG	150	Standard	58	Sanger	FALSmq2
PCDHB11_p.S759T_R	TCCAAAGCTATTTCGAAAGGTG					
PCNXL4_p.I576V_F	TAGTATGCTTACCCGCAGAGTG	171	Standard	58	Sanger	FALSmq2
PCNXL4_p.I576V_R	CTTTGTCCTTGGCAATTTCTTC					
FANCC_Ex7_F	TGTCCTTAATTATGCATGGCTC	304	Standard	64	Sanger	FALSmq2
FANCC_Ex7_R	TCGTACAGTCTTTCCAACACAC					
ANKRD18B_p.L589R_F	TAGAGGATGCTCGTAAGGAAGG	247	Standard	64	Sanger	FALSmq2
ANKRD18B_p.L589R_R	cgttcaacatatagccacaatga					
CHDC2_Ex3_pt1_F	aatacgtgaatgcatgttttgc	168	Standard	58	Sanger	FALSmq2
CHDC2_Ex3_pt1_R	GCATCATTGAATTTGCTAGTGG					
CHRNA2_M13FP_p.E411Q_F	CTGAAGCTCAGCCCCTCTTAT	245	Standard & PCR enhancer	61	Sanger	FALSmq2
CHRNA2_M13R-pUC_p.E411Q_R	ATGTAGTGCACACCTTCCAGTG					
ZFHX2_M13FP_p.T565S_F	GCTGTGACGTCTGCAACTACTC	225	Standard & PCR enhancer	58	Sanger	FALSmq2

ZFHX2_M13R-pUC_p.T565S_R	GGTTACGGGAGATGTTTGTCTC					
RASGRF1_Ex1_F	CGAGCGCCAGAGAGAGG	476	Standard	64	Sanger	FALSmq20
RASGRF1_Ex1_R	AGTAGAGGGGCCAAAGTTCAAG					
CRIM1_Ex17_F	caggctatcaatcaatgaattgtg	403	Standard	57	Sanger	FALSmq20
CRIM1_Ex17_R	TCCAATACAATCCACTAAGCAAG					
HR_Ex5_F	AGCAAGTGCGGAGgtgag	396	Standard	64	Sanger	FALSmq20
HR_Ex5_R	tcttctgaggttgccttaggtc					
TAZ_Ex1_F	AGTCAGGGGCCAGTGTCTC	279	Standard & PCR enhancer	66	Sanger	FALSmq20
TAZ_Ex1_R	ccaccctaagtcccacctc					
NCOR2_Ex43_F	CCCCAGAAGGTTCTGTCAGG	328	Standard & PCR enhancer	66	Sanger	FALSmq20
NCOR2_Ex43_R	CAGCTCTGAGGCAGGCAG					
DNAJC4_Ex4_F	ctctctggccttctgaggagt	413	Standard	64	Sanger	FALSmq20
DNAJC4_Ex4_R	agggacattgtggaaatgagac					
SOX15_p.R119Q_F	CAAGATGCACAACTCCGAGA	208	Standard	64	Sanger	FALSmq20
SOX15_p.R119Q_R	CAGGTTGCCTCTTCCCTGT					
HIC2_p.T526M_F	AGGAAGAGCTGTTCATCAAGGA	233	Standard	64	Sanger	FALSmq20
HIC2_p.T526M_R	CGTGAACATTTTGCCACAGAT					
ATP1B2_Ex7_F	ctagaccctgcactgctcctc	234	Standard	64	Sanger	FALSmq20
ATP1B2_Ex7_R	GAGCATCCACAGGAGAGAGATG					
POU2F2_Ex5_F	ACAGGTGGGCATTCTCTCTG	254	Standard	64	Sanger	FALSmq20
POU2F2_Ex5_R	CTAGCCCTGTAACAGATGAGGG					
MUC6_mq28_m13_F	GTACAGGAACCCCACCAATG	300	Standard	58	Sanger	FALSmq28
MUC6_mq28_m13_R	AGGATGTTGCAGTGACAGGAC					
GXYLT1_Ex5_mq28_m13_F	ACGACCAGTTGATGATATTTGG	186	Standard	62	Sanger	FALSmq28
GXYLT1_Ex5_mq28_m13_R	TTCCTTCTCATTCGAGTCATGTT					
GXYLT1_Ex1_mq28_m13_F	CTGCGGGTGAGGAACTTG	439	Standard + enhancer	64	Sanger	FALSmq28
GXYLT1_Ex1_mq28_m13_R	GAAAGACGCGGGAGACG					
MUC16_Ex55_mq28_m13_F	ccttctaccacacccctatgac	187	Standard	62	Sanger	FALSmq28
MUC16_Ex55_mq28_m13_R	agaagggaagcaggtcaaact					
MUC3A_Ex2_mq28_m13_F	CTACTTCTCCCACCAGCACTGT	236	Standard	65	Sanger	FALSmq28
MUC3A_Ex2_mq28_m13_R	ACTGTGTGAGGTGACTGTGGAG					
IQCE_mq28_m13_F	GTGTTGCATCTGGTGTCTGG	244	Standard	62	Sanger	FALSmq28
IQCE_mq28_m13_R	TCTAAGGCTCTTCCTCTTGACG					
HRNR_Ex3_mq28_m13_F	CAGGAGGGATCTAGCACAGG	150	Standard	62	Sanger	FALSmq28
HRNR_Ex3_mq28_m13_R	GCTGGAAGAGTGCCCAGA					
FRMPD2_Ex27_mq28_m13_F	aaatctaacagtgaatcttggtttttc	227	Standard	62	Sanger	FALSmq28
FRMPD2_Ex27_mq28_m13_R	cactgaacctatctacaaccctcttag					
PABPC3_Ex1_mq28_m13_F	ATGAAGATGCACAGAAAGCTGTAG	206	Standard	62	Sanger	FALSmq28
PABPC3_Ex1_mq28_m13_R	ACGTTCATCATCAATACCATCATC					
GAGE_gDNA_mq28_m13_F	CTTGACCTGCTGGGCTCAAGCG	3161	Standard	64	Sanger	FALSmq28
GAGE_gDNA_mq28_m13_R	ACCAGTCAAGGGTTCTTGGATA					
GAGE_Ex_mq28_m13_F	aaatatgagttggcgaggaaga	165	Standard	64	Sanger	FALSmq28
GAGE_Ex_mq28_m13_R	cacactaatgcaacgacgctat					
ZCCHC24_mq28_m13_F	TTCCATCTGCAGAATCACTTTG	437	Standard	64	Sanger	FALSmq28
ZCCHC24_mq28_m13_R	GACTGAAAGTGGGAGAGGTGAC					
TFCP2_mq28_m13_F	TTGTTGGGATTACAGGCATAAG	224	Standard	62	Sanger	FALSmq28
TFCP2_mq28_m13_R	GGAGGAGACGTATCTGGTTGTC					
TRIM69_mq28_m13_F	TGAGATCAGCCTAGGCAACATA	445	Standard	62	Sanger	FALSmq28
TRIM69_mq28_m13_R	CTGGCGTTTACACACTAGATGC					
FAM210A_mq28_m13_F	CTCCAGTCTGGGAGAGAGAGAG	242	Standard	62	Sanger	FALSmq28
FAM210A_mq28_m13_R	ATGGAGACTGCTAACTCCTTGG					
BAGE_mq28_m13_F	CAGTGGGAGAAGGGTAAAGAGA	176	Standard	62	Sanger	FALSmq28

BAGE_mq28_m13_R	TTCCTCTGGCCACACTTTCTAT					
RAB28_mq28_m13_F	GCTTTTGTTCCCACTTCAATGT	216	Standard	62	Sanger	FALSmq28
RAB28_mq28_m13_R	CCTGCCTTTCTACCTCTTCTCA					
OSTF1_mq28_m13_F	AAATGAATTGTTCCCACCTTTG	226	Standard	62	Sanger	FALSmq28
OSTF1_mq28_m13_R	CCCTTGAAAAATCCATTCAAAA					
SPRY3_mq28_m13_F	TGATATGGACTGCATCTGCTTT	262	Standard + enhancer	58	Sanger	FALSmq28
SPRY3_mq28_m13_R	GCTAAGCAGGAAACACCCTCTA					
DDX11L16_gDNA_mq28_m13_F	AACTACATGCAGGAACAGCAAA	224	Standard	62	Sanger	FALSmq28
DDX11L16_gDNA_mq28_m13_R	CCCACCAGCAATGTCTAGGAGT					
DNM1P41_mq28_m13_F	GTGGGCATGTGTGTGAGTGTGT	4000	Standard	64	Sanger	FALSmq28
DNM1P41_mq28_m13_R	GTATTTGCCAAGTTTTCTTAGA					
DNM1P41_mq28_m13_F	AAACACATCCCTCCTCTTCTCA	317	Standard	64	Sanger	FALSmq28
DNM1P41_mq28_m13_R	GACTGATTCCCAGTGCTAGAGG					
AKR1C2_mq28_F	GGGCAGGACATCGAAGATATCA	594	Standard	68	Sanger	FALSmq28
AKR1C2_mq28_R	AAACTTGCTGGGATGCCTATCA					
SLCO1C1_mq28_F	AGCTCTGTTTCTCTGCAACTGA	520	Standard	68	Sanger	FALSmq28
SLCO1C1_mq28_R	TCACTAGGGTGGTCTCTGTCTT					
CES1P1_mq28_F	GGGCTTTTCTGATCTCTCCCAA	536	Standard	68	Sanger	FALSmq28
CES1P1_mq28_R	CGCTATCCGTTATCGAGCCATA					
ERVV-2_mq28_F	TGACTTTGGAAAAGGAGGTGCT	565	Standard	68	Sanger	FALSmq28
ERVV-2_mq28_R	AGAGAAGTGCTGACTGTCTGTG					
MIR512_mq28_F	CAAACACACCCAGCTGAGTTTTAA	513	Standard	58	Sanger	FALSmq28
MIR512_mq28_R	CCCAGCCTGAATAACACCTTTTAC					
LINC01410_mq29_F	ACTAAGTGCATTCCTGGACCTG	564	Standard	68	Sanger	FALSmq28
LINC01410_mq29_R	TTGTCTGTCCTCTGCACATCTC					
MIR4477_mq30_F	CTCTGAAAATCTCAGGGCCCTT	623	Standard	68	Sanger	FALSmq28
MIR4477_mq30_R	TCTCCATCTAAGCTTCTGGGGA					
CES1P1_new_mq28_F	AGCTTGAAATCCTGGAACGCTA	536	Standard	63	Sanger	FALSmq28
CES1P1_new_mq28_R	CTTCAGAAGGACTCACCCCAAG					
MIR4477_new_mq28_FAM_F	FAM-CTGGGCAACAAGAGTGAAACTA	180	Standard	63	Fragment length analysis	FALSmq28
MIR4477_new_mq28_R	CATCTAAGCTTCTGGGGAGCTA					
PALM3_Ex6_p.E279G_F	AGACAGGAAGGGAGCTGGTAG	185	Standard	69	Sanger	Female SALS twins
PALM3_Ex6_p.E279G_R	AAGCTGCTGCCTCTAATCTCTC					
ZNF681_Ex4_p.K469K_F	TTCATACCAGAGAGAAACTCAATG	654	Standard	62	Sanger	Female SALS twins
ZNF681_Ex4_p.K469K_R	GTTTTGAGGATCGATAGAAAGC					
TYRO3_Ex17_p.V715delinsVWAFG_F	AAGGCTGACTCTCTCCCTCAAT	248	Standard	68	Sanger	Female SALS twins
TYRO3_Ex17_p.V715delinsVWAFG_R	TAAAGGTCTGCTCTCACACTCG					
ZNF571_Ex4_p.F447Y_F	GGGAAAGCCTTTATTTCTAATTCT	206	Standard	61	Sanger	Female SALS twins
ZNF571_Ex4_p.F447Y_R	TGTTGAGTAAGATATGCAACACGA					
KMT2C_Ex16_p.S902S_F	CAGAATGTTGACTTTTCCCAATC	220	Standard	63	Sanger	Female SALS twins
KMT2C_Ex16_p.S902S_R	AACTTGCTATGAGATTTTCATCATT					
CATSPER2_Ex11_p.E437fs_F	GGAGCTAGTCAACAAAGGGAAA	170	Standard	64	Sanger	Female SALS twins
CATSPER2_Ex11_p.E437fs_R	GAAGAGGATGTGGAGGAGACAC					
FNDC7_Ex6_p.L312L_F	AAGATGCGAAGAAAATCTCCTG	384	Standard	66	Sanger	Female SALS twins
FNDC7_Ex6_p.L312L_R	AATAGGTGATCCATTCTGCTGC					
ZNF718_Ex2_p.I11V_F	gataattccagtcagccccata	160	Standard	66	Sanger	Female SALS twins
ZNF718_Ex2_p.I11V_R	gGTCCAGACATTTCCACTCTTC					
ZNF771_Ex3_p.T86T_F	cgctaagggctgacctatcc	213	Standard + enhancer	61	Sanger	Female SALS twins
ZNF771_Ex3_p.T86T_R	GTCAGCGCCGACTTCTGT					
FLG2_Twin set set2_F	GAATCCATAGTTCCTGAGAGACATG	527	Standard	66	Sanger	*SOD1* triplets
FLG2_Twin set set2_R	GATTGAGAATGTCCACTGGTATCTC					
PSPC1_N450S_F	CTCCTCCAATGATGGGTATGAA	236	Standard	61	Sanger	*SOD1* triplets

PSPC1_N450S_R	TGCCAAAACAAATGATACCAAA					
POTEB2_E249Q_F	AGAAGGAAGCGACCAAATTTTA	395	Standard	64	Sanger	*SOD1* triplets
POTEB2_E249Q_R	CTTTATGTTGCCCAGTCCAAAT					
PRAMEF36P_SOD1triplets_F	CTCACTGTCATCATTGGTCCAT	176	Standard	67	Sanger	*SOD1* triplets
PRAMEF36P_SOD1triplets_R	ACATGCAAATTCAAGGCTAGGT					
LOC390705_SOD1triplets_F	ATATTTGGGATCCTGCTGAGG	212	Standard	64	Sanger	*SOD1* triplets
LOC390705_SOD1triplets_R	TATTTCCTTTTCCACCATACGC					
MIR4436A_SOD1triplets_new_F	FAM-GATGGATCTTTCCCAACTTCTG	323	Standard	63	Fragment length analysis	*SOD1* triplets
MIR4436A_SOD1triplets_new_R	AAATTTCAAACCCCCAAAAAGT					
SMOC2_SOD1triplets_new_F	FAM-AGGCTTGCCTATAAATGAGGTG	673	Standard	61	Fragment length analysis	*SOD1* triplets
SMOC2_SOD1triplets_new_R	CCTATTCAGGCCAACCTACAAC					
OR1L3_SOD1triplets_new_F	FAM-GCTACAACCTCCACCTCCCACC	293	Standard	71	Fragment length analysis	*SOD1* triplets
OR1L3_SOD1triplets_new_R	CGGGGCGGCTGGCTGGGCAGAGG					
OR2T2_V320fs_F	CAACCCACTCATCTACAGCTTG		Standard	68	Sanger	*C9orf72* twins
OR2T2_V320fs_R	GTTTGCGCTAGTCCTTGCTAGT					
C5orf60_K100N_F	CCATCTCCTGCTTTCCAGATTA		Standard	64	Sanger	*C9orf72* twins
C5orf60_K100N_R	ACTTCACAAAGCTCTGCCTACC					
VPS52_87_88del_F	CCCCAAAAGGAAAAACAACA		Standard	61	Sanger	*C9orf72* twins
VPS52_87_88del_R	CCCTCTGCTGGAGGATACAA					
SPATA31C1_M154T_F	GACTTTGGTCAGCTCTCTGGTC		Standard	68	Sanger	*C9orf72* twins
SPATA31C1_M154T_R	AGTTGAAGATGCACCAGGTTTT					
AHNAK_2384_2386del_F	GATGCTGACATGCCAGAAGTAG		Standard	64	Sanger	*C9orf72* twins
AHNAK_2384_2386del_R	CCTTCCAATTTGGGAACATCTA					
PHC1_Q434delinsQQQ_F	TGGACGAAGTGATGTCCAAG		Standard	68	Sanger	*C9orf72* twins
PHC1_Q434delinsQQQ_R	AGTAGGTGGGACCTGTGGTG					
SCAF1_S827F_F	ACAGGGACAGAGATAGGGACAG		Standard	66	Sanger	*C9orf72* twins
SCAF1_S827F_R	GACCTTTTTCCTGGTCTTGGAT					

*Touchdown PCR: thermocycling conditions in which the annealing temperature decreases 2 degrees each cycle, after the first 10 cycles.
Standard: standard PCR reaction conditions included MyTaq HS Red mix, milliQ water, 10mM forward and reverse primers and 20ng DNA.

A.3.2

A.3.3 ACMG guidelines for interpreting sequence variants

The following table outlines the ACMG recommended guideline for interpreting the pathogenic nature of sequencing variants. These guiudelines were consulted as part of the development of the *in silico* assessment of pathogenicity pipeline presented in Chapter 6. Obtained from Richards et al. (2015).

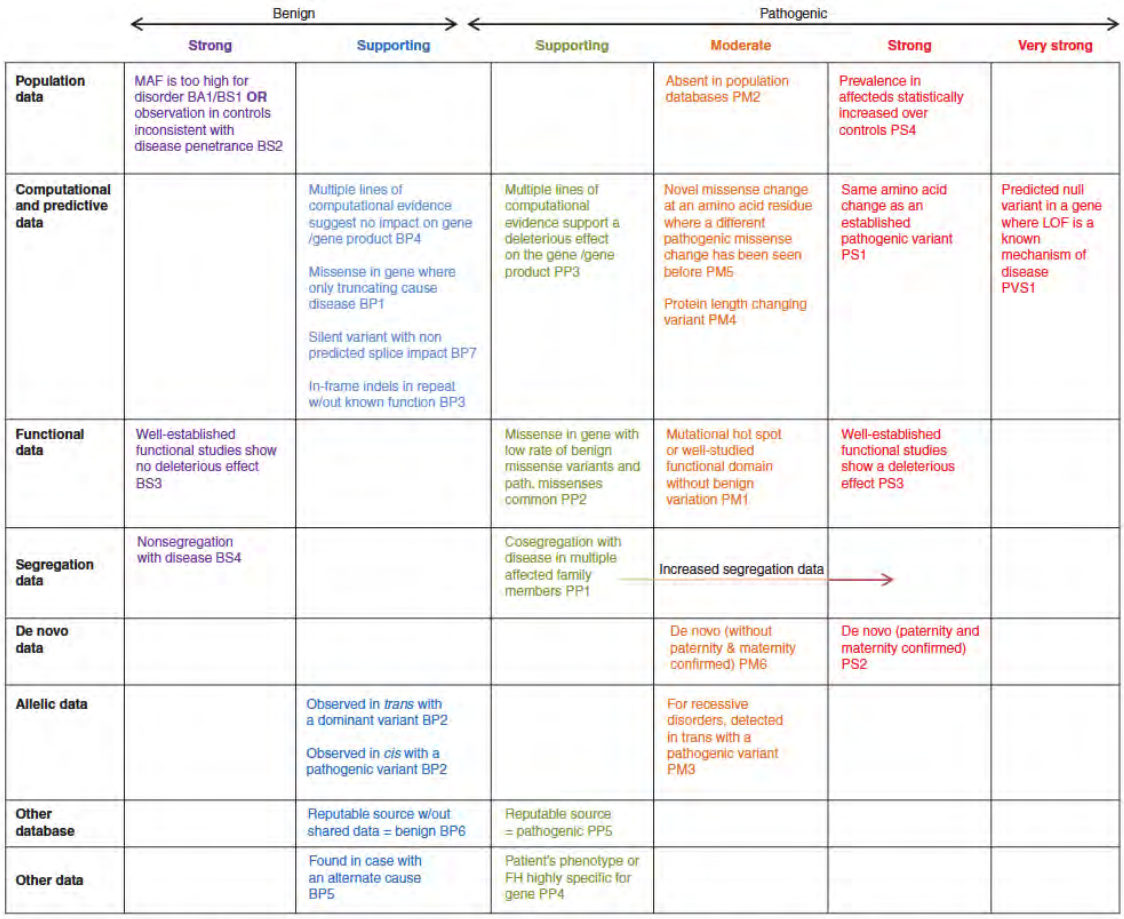


Figure 1 Evidence framework. This chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. Evidence code descriptions can be found in **Tables 3** and **4**. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

A.3.3.1 FALS-associated candidate gene variants

The following table provides details of all population-based SNPs found to be associated with FALS compared with ExAC control individuals. Details include allele counts in the patient and control cohorts, and the p-value results of Fisher's exact testing comparing patients to the various control cohorts. The two SNPs which are highlighted represent those which withstood family bias testing after applying Bonferroni correction to the p-value threshold, however both subsequently failed replication using Australian control cohorts.

Gene	CHROM	POS	rs ID	FALS patients			ExAC control comparison					Diamantina control comparison					MGRB control comparison					Potential disease-risk or protective allele?	Association conclusion	Notes
				MAF	Ref AC	Alt AC	MAF	Ref AC	Alt AC	Fisher's p-value	Family biased result?	MAF	Ref AC	Alt AC	Fisher's p-value	Family biased result?	MAF	Ref AC	Alt AC	Fisher's p-value	Family biased result?			
SCFD1	14	31099738	rs229150	0.346	106	56	0.433	68857	52493	2.629E-02	yes										protective	family biased	
SPTBN4	19	40978640	rs73931308	0.092	129	13	0.206	65606	17054	3.661E-04	no	0.151	1446	258	6.305E-02		0.156	1932	356	4.004E-02	yes	protective	family and/or population biased	
SPTBN4	19	41018832	rs814533	0.075	124	10	0.176	4211	901	1.137E-03	no	0.000	1920	0	1.010E-12	no	0.106	2035	241	3.080E-01		protective	conflicting population results	absent from one Aust. Control cohort
SPTBN4	19	41060616	rs2242131	0.082	134	12	0.217	20394	5654	1.890E-05	no	0.149	1036	182	3.213E-02	no	0.170	1899	389	3.884E-03	no	protective	disease associated	
SPTBN4	19	41071552	.	0.027	144	4	0.003	18827	51	8.812E-04	no	0.208	1531	401	5.160E-10	no	0.000	2288	0	1.310E-05	no	risk	disease associated	absent from one Aust. Control cohort
C21orf2	21	45750145	rs11552066	0.025	119	3	0.156	53959	9991	4.060E-06	no	0.112	215	27	3.991E-03	no	0.120	2014	274	3.761E-04	no	protective	disease associated	
C21orf2	21	45750346	rs2070573	0.130	40	6	0.347	10719	5707	1.598E-03	no	0.206	54	14	3.288E-01		0.219	1786	502	2.046E-01		protective	population biased	
C21orf2	21	45759045	rs11870	0.091	120	12	0.342	9451	4905	2.990E-11	no	0.500	4	4	6.193E-03	no	0.222	1781	507	1.784E-04	no	protective	disease associated	
NEK1	4	170428331	rs6855803	0.203	126	32	0.308	17749	7917	4.158E-03	yes											protective	family biased	
NEK1	4	170506525	rs200161705	0.025	158	4	0.002	0	0	6.207E-04	no	0.004	1926	8	1.041E-02	no	0.003	2280	8	5.996E-03	no	risk	disease associated	
DDX58	9	32481339	rs61752945	0.037	156	6	0.009	119973	1079	3.566E-03	yes											risk	family biased	
DDX58	9	32500832	rs72710678	0.031	157	5	0.009	120167	1137	1.904E-02	no	0.024	1887	47	5.948E-01		0.020	2242	46	3.832E-01	yes	risk	population biased	
EEF1A2	20	62124459	rs12480745	0.133	104	16	0.258	12443	4325	1.516E-03	no	0.312	919	417	1.600E-05	no	0.267	1676	612	8.529E-04	no	protective	disease associated	
EEF1A2	20	62126185	rs310617	0.667	46	92	0.571	50722	67506	2.505E-02	yes											risk	family biased	
EEF1A1	6	74227940	rs11556677	0.080	149	13	0.170	0	0	1.545E-03	no	0.238	1474	460	6.770E-07	no	0.000	2288	0	2.910E-16	no	protective	disease associated	absent from one Aust. Control cohort
EEF1D	8	144662353	rs1062391	0.615	57	91	0.511	58399	61119	1.336E-02	no	0.626	724	1210	7.922E-01		0.653	793	1495	3.736E-01	yes	risk	population biased	
EEF1D	8	144671922	rs3812448	0.836	24	122	0.760	26221	83105	3.267E-02	yes											risk	family biased	
ANXA11	10	81917463	.	0.006	159	1	0.000	121234	4	6.573E-03	no	0.000	1920	0	7.692E-02		0.000	2288	0	6.536E-02		risk	population biased	absent from both Aust. Control cohort
ANXA11	10	81921715	rs146222704	0.014	146	2	0.000	119952	34	9.239E-04	no	0.001	1931	1	1.439E-02	no	0.000	2288	0	3.668E-03	yes	risk	conflicting population family results	absent from one Aust. Control cohort
ANXA11	10	81921810	rs2304410	0.092	138	14	0.155	94298	17280	3.243E-02	yes											protective	family biased	
ANXA11	10	81926702	rs1049550	0.434	86	66	0.309	0	0	1.119E-03	no	0.427	1109	825	8.651E-01		0.425	1315	973	8.656E-01		risk	population biased	
ANXA11	10	81926718	rs2228427	0.092	138	14	0.177	70596	15140	5.307E-03	yes											protective	family biased	
ANXA11	10	81926750	rs34332933	0.092	138	14	0.175	52786	11178	5.279E-03	yes											protective	family biased	
ANXA11	10	81930787	.	0.008	131	1	0.000	11196	2	3.455E-02	yes											risk	family biased	
GGNBP2	17	34935878	rs2074103	0.519	78	84	0.392	73526	47394	1.222E-03	yes											risk	family biased	
GGNBP2	17	34941864	.	0.006	161	1	0.000	121156	2	4.001E-03	no	0.000	1920	0	7.781E-02		0.000	2288	0	6.612E-02		risk	population biased	absent from both Aust. Control cohort
GGNBP2	17	34942595	rs1106908	0.519	78	84	0.392	72963	47035	1.222E-03	yes											risk	family biased	
GGNBP2	17	34943719	rs3744593	0.534	68	78	0.397	72342	47700	9.166E-04	yes											risk	family biased	
GPX3	5	150406437	rs869976	0.006	161	1	0.042	113232	4994	1.654E-02	no	0.010	1915	19	1.000E+00		0.007	2272	16	1.000E+00		protective	population biased	absent from both Aust. Control cohort
GPX3	5	150407456	rs8177447	0.736	38	106	0.832	0	0	3.517E-03	yes											protective	family biased	
TNIP1	5	150439921	rs200236985	0.006	161	1	0.000	121341	17	2.373E-02	no	0.000	1920	0	7.781E-02		0.001	2286	2	1.856E-01	yes	risk	population biased	absent from one Aust. Control cohort
TNIP1	5	150443266	.	0.006	157	1	0.000	121362	12	1.677E-02	no	0.000	1920	0	7.603E-02		0.000	2287	1	1.250E-01	yes	risk	population biased	absent from one Aust. Control cohort
ABCC2	10	101569997	rs17222639	0.074	150	12	0.031	113618	3648	5.050E-03	yes											risk	family biased	
ABCC2	10	101577125	.	0.006	161	1	0.000	121344	2	3.994E-03	no	0.000	1920	0	7.781E-02		0.000	2288	0	6.612E-02		risk	population biased	absent from both Aust. Control cohort
ABCC2	10	101590619	rs41318031	0.006	161	1	0.036	115388	4270	3.318E-02	yes											protective	family biased	
ABCC2	10	101604207	rs3740066	0.420	94	68	0.342	0	0	3.857E-02	yes											risk	family biased	
UBA1	X	47062534	rs2070169	0.117	143	19	0.183	71538	16075	3.216E-02	yes											protective	family biased	
MTHFR	1	11848139	rs11559040	0.149	126	22	0.237	9151	2839	1.107E-02	yes											protective	family biased	
BICD2	9	95480120	rs142140690	0.006	157	1	0.000	120163	7	1.046E-02	no	0.000	1920	0	7.603E-02		0.000	2288	0	6.460E-02		risk	population biased	absent from both Aust. Control cohort
BICD2	9	95483066	.	0.013	158	2	0.000	113773	37	1.403E-03	no	0.001	1933	1	1.654E-02	no	0.001	2285	3	3.724E-02	no	risk	disease associated	
BICD2	9	95526977	.	0.034	141	5	0.003	46235	161	1.869E-04	no	0.000	1920	0	1.650E-06	no	0.000	2288	0	7.280E-07	no	risk	disease associated	absent from both Aust. Control cohort
CHCHD1	10	75542163	rs139732935	0.006	157	1	0.000	121177	1	2.603E-03	no	0.000	1920	0	7.603E-02	0	0.000	2288	0	6.460E-02	yes	risk	population biased	absent from one Aust. Control cohort
CHCHD5	2	113346480	rs41278942	0.160	126	24	0.097	107661	11623	1.795E-02	yes											risk	family biased	
CHCHD6	3	126676314	rs145020754	0.014	146	2	0.002	76642	168	4.278E-02	no	0.002	1893	3	4.504E-02	no	0.001	2285	3	3.247E-02	no	risk	disease associated	
PINK1	1	20960230	rs45530340	0.113	126	16	0.225	9876	2868	7.904E-04	no	0.000	1920	0	1.130E-19	no	0.200	1827	457	8.734E-03	yes	protective	conflicting population family results	absent from one Aust. Control cohort
PINK1	1	20977000	rs1043424	0.383	100	62	0.297	85221	36071	2.011E-02	no	0.266	1418	514	1.823E-03	yes	0.276	1656	632	5.003E-03	no	risk	conflicting population family results	
PINK1	1	20977107	.	0.006	161	1	0.000	121325	3	5.323E-03	no	0.000	1920	0	7.781E-02		0.000	2288	0	6.612E-02		risk	population biased	absent from both Aust. Control cohort
CNR2	1	24200983	rs2229583	0.481	84	78	0.620	43846	71560	3.489E-04	no	0.566	838	1094	3.948E-02	no	0.583	955	1333	1.347E-02	yes	protective	conflicting population family results	
CNR2	1	24201109	rs2229580	0.481	84	78	0.619	46251	75147	4.626E-04	no	0.567	836	1096	3.920E-02	no	0.583	955	1333	1.347E-02	yes	protective	conflicting population family results	
CNR2	1	24201262	rs2502993	0.459	80	68	0.619	46225	75111	8.870E-05	no	0.569	833	1099	1.247E-02	no	0.583	955	1333	4.490E-03	yes	protective	conflicting population family results	
CNR2	1	24201357	rs4649124	0.447	84	68	0.618	46187	74839	2.480E-05	no	0.568	835	1097	4.979E-03	no	0.583	955	1333	1.252E-03	no	protective	disease associated	
CNR2	1	24201448	rs3003336	0.481	81	75	0.619	45725	74331	4.976E-04	no	0.570	830	1102	3.563E-02	no	0.583	955	1333	1.499E-02	yes	protective	conflicting population family results	
CNR2	1	24201643	rs2501431	0.475	83	75	0.620	45979	74967	2.805E-04	no	0.569	832	1100	2.416E-02	no	0.583	955	1333	9.706E-03	yes	protective	conflicting population family results	
CNR2	1	24201919	rs2502992	0.449	86	70	0.618	46161	74775	2.190E-05	no	0.565	841	1091	5.632E-03	no	0.583	955	1333	1.406E-03	no	protective	disease associated	
CNR2	1	24201920	rs2501432	0.449	86	70	0.618	46192	74778	2.200E-05	no	0.565	840	1092	5.597E-03	no	0.583	955	1333	1.406E-03	no	protective	disease associated	
FAAH	1	46860237	rs72890727	0.254	106	36	0.350	480	258	2.581E-02	yes											protective	family biased	
FAAH	1	46871039	rs200731801	0.013	152	2	0.002	120942	202	2.806E-02	yes											risk	family biased	
FAAH	1	46871409	rs114795065	0.013	148	2	0.001	119485	111	9.005E-03	no	0.004	1902	8	1.614E-01		0.002	2284	4	4.789E-02	no	risk	conflicting population results	
FAAH	1	46871746	rs41305628	0.066	142	10	0.032	117391	3909	3.302E-02	yes											risk	family biased	
DAGLA	11	61496352	.	0.006	153	1	0.000	119903	3	5.121E-03	no	0.000	1920	0	7.425E-02		0.000	2288	0	6.306E-02		risk	population biased	absent from both Aust. Control cohort
DAGLA	11	61504644	.	0.007	147	1	0.000	121298	2	3.651E-03	yes											risk	family biased	
DAGLA	11	61507041	.	0.014	146	2	0.000	120850	2	8.900E-06	no	0.000	1920	0	5.090E-03	no	0.000	2288	0	3.668E-03	no	risk	disease associated	absent from both Aust. Control cohort
MGLL	3	127540635	rs11538698	0.074	150	12	0.031	117068	3684	4.342E-03	no	0.047	1843	91	1.296E-01		0.032	2214	74	1.216E-02	yes	risk	population and/or family biased	
CNR1	6	88853635	rs1049353	0.297	111	47	0.214	95470	25926	1.472E-02	yes											risk	family biased	
DAGLB	7	6449794	rs139753251	0.027	146	4	0.008	77435	649	3.762E-02	yes											risk	family biased	
DAGLB	7	6456319	rs1055430	0.095	134	14	0.042	114862	5032	5.732E-03	no	0.062	1773	117	1.182E-01		0.061	2149	139	1.131E-01		risk	population biased	
DAGLB	7	6461348	rs836514	0.101	133	15	0.175	94932	20156	1.684E-02	yes											protective	family biased	
DAGLB	7	6461424	rs836515	0.006	157	1	0.035	116199	4165	4.753E-02	yes											protective	family biased	
NAPEPLD	7	102743893	rs3181009	0.994	1	157	0.959	4913	116259	2.355E-02	no	0.999	1	1927	1.458E-01		0.998	5	2283	3.304E-01		risk	population biased	
KCND3	1	112524680	rs17215423	0.047	141	7	0.014	0	0	5.119E-03	yes											risk	family biased	
CDT1	16	88872122	.	0.014	146	2	0.000	0	0	3.850E-04	no	0.000	1920	0	5.090E-03	no	0.000	2288	0	3.668E-03	yes	risk	conflicting population family results	absent from both Aust. Control cohort
CDT1	16	88872176	rs145319641	0.007	149	1	0.000	0	0	8.731E-03	no	0.000	1920	0	7.246E-02		0.000	2288	0	6.153E-02		risk	population biased	absent from both Aust. Control cohort
CDT1	16	88872511	rs510862	0.750	37	111	0.818	18771	84557	4.182E-02	no	0.774	407	1395	5.405E-01		0.785	493	1795	3.545E-01		protective	population biased	
CDT1	16	88872542	rs3218722	0.041	142	6	0.130	47922	7130	4.958E-04	no	0.045	1713	81	1.000E+00		0.059	2152	136	4.673E-01	yes	protective	population biased	absent from one Aust. Control cohort
CDT1	16	88873729	.	0.007	147	1	0.000	116192	20	2.638E-02	no	0.000	1920	0	7.157E-02		0.000	2288	0	6.076E-02		risk	population biased	absent from both Aust. Control cohort
CDT1	16	88873738	rs200199040	0.007	147	1	0.000	116434	10	1.388E-02	no	0.000	1920	0	7.157E-02		0.000	2288	0	6.076E-02		risk	population biased	absent from both Aust. Control cohort
CDT1	16	88874632	rs572275	0.349	95	51	0.474	0	0	2.741E-03	no	0.346	1207	639	9.283E-01		0.352	1482	806	1.000E+00		protective	population biased	absent from one Aust. Control cohort
KIF5A	12	57963020	rs181688415	0.037	156	6	0.008	120111	981	2.240E-03	no	0.017	1902	32	6.785E-02	no	0.014	2256	32	3.596E-02	no	risk	disease associated	
KIF5A	12	57969016	rs144382702	0.006	157	1	0.000	112105	17	2.503E-02	no	0.000	1920	0	7.603E-02		0.000	2288	0	6.460E-02		risk	population biased	absent from both Aust. Control cohort

A.3.3.2 SALS-associated candidate gene variants

The following table provides details of all population-based SNPs found to be associated with SALS compared with gnomAD NFE control individuals. Details include allele counts in the patient and control cohorts, and the p-value results of Fisher's exact testing comparing patients to the various control cohorts.

Gene	CHROM	POS	rs ID	Gene function	SALS patients			gnomAD NFE control comparison				Diamantina control comparison				MGRB control comparison				Potential disease-risk or protective allele?	Association conclusion
					MAF	Ref AC	Alt AC	MAF	Ref AC	Alt AC	Fisher's p-value	MAF	Ref AC	Alt AC	Fisher's p-value	MAF	Ref AC	Alt AC	Fisher's p-value		
CHCHD2	7	56171871	rs374406633	intronic	0.001	1255	1	0.000	123318	4	4.940E-02	NA	NA	NA	NA	0.000	2288	0	3.544E-01	risk	population biased
CHCHD3	7	132481193	.	intronic	0.001	1255	1	0.000	125680	0	9.895E-03	NA	NA	NA	NA	0.000	2288	0	3.544E-01	risk	population biased
CHCHD3	7	132570458	.	exonic	0.001	1255	1	0.000	125691	3	3.899E-02	0.000	1920	0	3.955E-01	0.000	2288	0	3.544E-01	risk	population biased
CHCHD3	7	132719349	rs78193687	intronic	0.059	1182	74	0.075	112609	9093	3.501E-02	NA	NA	NA	NA	0.079	2107	181	2.948E-02	protective	disease associated
CHCHD3	7	132719440	.	intronic	0.001	1255	1	0.000	125927	3	3.892E-02	NA	NA	NA	NA	0.000	2288	0	3.544E-01	risk	population biased
CHCHD6	3	126423125	rs200230339	UTR5	0.001	1253	1	0.000	117353	1	2.103E-02	NA	NA	NA	NA	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126423136	rs200360858	UTR5	0.001	1253	1	0.000	116601	1	2.117E-02	NA	NA	NA	NA	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126445926	.	exonic	0.001	1253	1	0.000	126576	0	9.810E-03	0.000	1920	0	3.951E-01	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126449513	.	intronic	0.001	1253	1	0.000	123539	1	2.000E-02	NA	NA	NA	NA	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126633636	rs199708316	intronic	0.001	1253	1	0.000	121474	0	1.022E-02	NA	NA	NA	NA	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126676329	rs77373684	exonic	0.001	1253	1	0.000	117544	0	1.056E-02	0.000	1920	0	3.951E-01	0.000	2288	0	3.540E-01	risk	population biased
CHCHD6	3	126676337	.	exonic	0.001	1253	1	0.000	113086	0	1.097E-02	0.000	1920	0	3.951E-01	0.000	2288	0	3.540E-01	risk	population biased
TIA1	2	70451784	rs76438450	intronic	0.002	1253	3	0.000	117011	3	2.330E-05	NA	NA	NA	NA	0.004	2279	9	5.567E-01	risk	population biased
TIA1	2	70463168	.	intronic	0.001	1255	1	0.038	60623	2395	1.410E-19	NA	NA	NA	NA	0.000	2288	0	3.544E-01	protective	population biased
TIA1	2	70463334	rs78928004	intronic	0.002	1254	2	0.000	48398	8	2.513E-02	NA	NA	NA	NA	0.019	2245	43	1.020E-06	risk	disease associated
TIA1	2	70475680	rs141564047	UTR5	0.002	1254	2	0.247	45939	15095	1.130E-148	NA	NA	NA	NA	0.000	2288	0	1.255E-01	protective	population biased

A.3.3.3 Replication of Fisher's exact testing using Project MiNE cohort

The following table provides details and results of Fisher's Exact testing in the Project MiNE case-control cohort for those variants found to be associated with disease in Australian FALS or SALS (Table 5.3).

Table A.2: **Replication of association testing in the Project MiNE case-control cohort.**

Gene	CHROM	POS	rsID	Project MiNE cases		Project MiNE control		Fisher's Exact
				Ref allele count	Alt allele count	Ref allele count	Alt allele count	p-vlaue
SPTBN4	19	41060616	rs2242131	0	0	0	0	no data available
SPTBN4	19	41071552	.	0	0	0	0	no data available
C21orf2	21	45750145	rs11552066	7729	1003	3251	413	7.57024E-01
C21orf2	21	45759045	rs11870	6903	1827	2863	801	2.47828E-01
NEK1	4	170506525	rs200161705	8664	68	3651	13	6.84479E-03
EEF1A2	20	62124459	rs12480745	6257	2471	2651	1013	4.56730E-01
EEF1A1	6	74227940	rs11556677	0	0	0	0	no data available
BICD2	9	95483066	.	8723	9	3658	6	3.99739E-01
BICD2	9	95526977	.	8706	0	3655	1	2.95745E-01
CHCHD6	3	126676314	rs145020754	8719	13	3651	11	1.13736E-01
CNR2	1	24201357	rs4649124	0	0	0	0	no data available
CNR2	1	24201919	rs2502992	0	0	0	0	no data available
CNR2	1	24201920	rs2501432	4936	3792	2195	1467	5.28027E-04
DAGLA	11	61507041	.	0	0	0	0	no data available
KIF5A	12	57963020	rs181688415	8634	98	3608	56	7.50365E-02
CHCHD3	7	132719349	rs78193687	8080	650	3377	287	4.56661E-01
TIA1	2	70463334	rs78928004	8725	7	3660	4	7.41571E-01

A.3.4 *In silico* assessment of pathogenicity results

The following tables describe the results obtained from application of the *in silico* pipeline for assessment of potential ALS pathogenicity.

Table A.3: ***In silico*** **assessment of pathogenicity results - proof of principle.**

Variant type	**Gene**	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**
			Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI and ClustalOmega	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score	
Known ALS mutation	*SOD1*	p.I114T	9	110.245	2	8/8	2	12/16	7.42246	1	1.8	-0.08 (47.79%)	2.34	2.2142	8.0142
Known ALS mutation	*SOD1*	p.E101G	9	110.245	2	5/7	1	7/16	0.871835	0	0.5	-0.08 (47.79%)	2.34	2.2142	5.7142
Known ALS mutation	*SOD1*	p.V149G	9	110.245	2	7/7	2	14/16	7.39868	1	1.8	-0.08 (47.79%)	2.34	2.2142	8.0142
Known ALS mutation	*FUS*	p.R521C	11.5	56.582	2	5/7	1	4/4	2.6095	1	1.6	-1 (8.32%)	2.6	3.1336	7.7336
Known ALS mutation	*FUS*	p.R521H	11.5	56.582	2	3/7	0.5	4/4	4.07511	1	1.6	-1 (8.32%)	2.6	3.1336	7.2336
Known ALS mutation	*FUS*	p.R521S	11.5	56.582	2	5/7	1	4/4	2.6095	1	1.6	-1 (8.32%)	2.6	3.1336	7.7336
Known ALS mutation	*TARDBP*	p.G294V	10	30.101	2	2/7	0	5/6	4.56656	1	1.6	-0.38 (27.42%)	4.33	3.6166	7.2166
Known ALS mutation	*TARDBP*	p.M337V	10	30.101	2	3/7	0.5	6/6	8.91094	1	1.8	-0.38 (27.42%)	4.33	3.6166	7.9166
Known ALS mutation	*TARDBP*	p.G376D	10	30.101	2	2/7	0	4/6	6.85943	1	1.4	-0.38 (27.42%)	4.33	3.6166	7.0166
Known ALS mutation	*UBQLN2*	p.T487I	10.5	13.089	2	0/7	0	3/6	1.65513	0.913386	1.4	-0.36 (28.63%)	1.56	2.2074	5.6074
Known ALS mutation	*CCNF*	p.S621G	5.5	0.401	0.5	4/7	0.5	5/5	6.86126	1	1.8	-1.22 (5.67%)	0.22	1.9966	4.7966
Common SNP	*TMA16*	p.I176T	no result	10.44	1	1/7	0	2/6	-0.498984	0	0	0.7 (85.42%)	-0.27	0.1566	1.1566
Common SNP	*OR4C3*	p.S100F	6	0	0.5	4/7	0.5	2/2	2.06728	0	1.1	-0.07 (48.69%)	-5.59	0	2.1
Common SNP	*MAP2K3*	p.S39P	5.5	7.65	0.5	2/7	0	3/5	1.876	0	0.9	0.11 (61.91%)	-0.23	0.6468	2.0468

Table A.4: ***In silico*** **assessment of pathogenicity results - proband candidate mutations.**

Gene	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**	**Priority category**
		Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI alignment	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score		
SPTBN4	p.R2074P	6	1.773	0.5	2 of 7	0	3 of 3	1	0.11811	1.1	60.71	0.85	1.9574	3.5574	medium
EEF1D	p.F278L	5	24.12	1	5 of 7	1	7 of 7	4.20623	1	2	2.13	absent	0.3154	4.3154	medium
ABCC2	p.D942N	4	0.097	0	0 of 8	0	2 of 6	0.279496	0	0.5	73.73	-0.42	0	0.5	low
ABCC2	p.N1186K	4	0.097	0	6 of 7	2	5 of 6	-0.0699685	0.755906	1.1	90.09	-2.48	0	3.1	medium
MTHFR	p.V541L	6.5	2.151	0.5	3 of 7	0.5	7 of 10	7.15759	1	1.6	90.09	-2.48	1.2108	3.8108	medium
DAGLB	p.E506K	8	4.649	1	6 of 7	2	4 of 6	5	1	1.4	55.86	1.51	1.6378	6.0378	high
TIA1	p.A254G	9	33.9	2	7 of 7	2	5 of 7	9.422	1	1.6	69.21	1.83	1.5308	7.1308	high
TIA1	p.P294L	9	33.9	2	3 of 7	0.5	3 of 7	6.88494	1	1	69.21	1.83	1.5308	5.0308	high
TIA1	p.H54N	9	33.9	2	2 of 7	0	3 of 7	7.1629	1	1	69.21	1.83	1.5308	4.5308	medium

Table A.5: ***In silico*** **assessment of pathogenicity results - FALS15 candidate mutations.**

Gene	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**	**Priority category**	**Priority Ranking**
		Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI alignment	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score			
CLCN4	p.I668T	9.5	11.236	2	5/7	1	4/9	7.82206	1	1	-1.09 (7.05%)	4.7	4	8	high	1
MTSS1L	p.A126T	10.5	93.487	2	5/8	1	9/9	5.7758	1	2	-1.55 (3.27%)	-0.16	1.8546	6.8546	high	2
SCN4A	p.R225W	4.5	0	0	8/8	2	9/9	5.21017	1	2	-0.75 (13.68%)	1.23	2.3414	6.3414	high	3
LRRN2	p.I196T	7	0.848	0.5	7/8	2	2/3	9.32553	1	1.2	-0.28 (33.53%)	2.11	2.3844	6.0844	high	6
SUPV3L1	p.Q168E	8.5	4.718	1	3/7	0.5	10/13	7.72159	1	1.8	-0.98 (8.85%)	1.12	2.383	5.683	high	5
HOXD3	p.Y249C	3.5	1.391	0	6/8	1	8/8	7.41859	1	2	0.11 (61.73%)	1.96	1.7454	4.7454	medium	7
FAM171A1	p.H518R	N/A	56.38	1	1/8	0	6/9	3.02023	1	1.2	-1.32 (4.73%)	0.88	2.3454	4.5454	medium	4
SP1	p.A145T	7	5.094	1.5	0/8	0	2/8	0.367055	0.0393701	0.5	-0.93 (9.55%)	0.56	2.089	4.089	medium	11
MAPKAPK3	p.K368R	6.5	3.804	0.5	2/7	0	10/12	6.67572	1	1.8	-0.25 (35.75%)	0.79	1.68	3.98	medium	8
SIM1	p.G733V	4.5	0	0	2/7	0	9/9	3.81394	1	2	-0.82 (11.88%)	0.15	1.8374	3.8374	medium	9
ZNF385B	p.P435S	7.5	0.318	0.5	3/7	0.5	8/9	3.21503	1	1.8	0.11 (61.73%)	0.2	0.8654	3.6654	medium	12
TYMP	p.Q245E	5.5	9.537	0.5	1/7	0	4/5	3.61484	1	1.6	no result	1.67	0.835	2.935	medium	10
TNS2	p.S992L	7.5	28.767	1.5	2/7	0	no data	3.75447	0.661417	1	no result	no result	0	2.5	medium	15
NECAB3	p.R203L	8.5	4.44	1	1/7	0	4/7	0.301457	0.00787402	1.1	no result	0.77	0.385	2.485	medium	13
ZNF425	p.R424P	5	0.871	0	3/6	0.5	2/2	-1.26231	0	0.6	-0.33 (30.86%)	-1.16	0.8028	1.9028	low	16
CEP295	p.N1707I	N/A	1.349	0	4/6	0.5	no data	1.44461	0.299213	0.5	no result	no result	0	1	low	14
ZNF497	p.K23R	5	1.01	0	1/8	0	1/1	-0.276772	0	0.6	no result	no result	0	0.6	low	17
ZNF497	p.V22G	5	1.01	0	1/8	0	1/1	-1.66251	0	0.6	no result	no result	0	0.6	low	18
RNF133	p.R94Q	3.5	0	0	3/8	0	2/6	-0.879315	0	0	0.75 (86.65%)	-1.7	0	0	low	19

Table A.6: ***In silico*** **assessment of pathogenicity results - FALS45 candidate mutations.**

Gene	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**	**Priority category**	**Priority Ranking**
		Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI alignment	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score			
SCCPDH	p.V256L	9	64.063	2	6/7	2	9/10	8.086	1	1.8	0.26 (70.44%)	0.62	0.9012	6.7012	high	1
GDPD1	p.P221T	7	1.627	0.5	6/7	2	5/5	7.18455	1	1.8	-0.19 (39.68%)	1.32	1.8664	6.1664	high	2
SPATA2	p.G206S	7.5	11.257	1.5	4/7	0.5	5/5	3.29511	0.992126	1.8	-0.86 (10.85%)	0.39	1.978	5.778	high	3
KRT85	p.S5P	4.5	0	0	6/7	2	4/4	0.714032	0.779528	1.6	-0.46 (23.63%)	1.3	2.1774	5.7774	high	4
GABRG3	p.S236F	9	0.023	1	3/7	0.5	6/7	5.50611	1	1.8	-0.09 (46.74%)	1.92	2.0252	5.3252	high	5
GRIN2D	p.V144L	5.5	0.382	0	3/7	0.5	2/4	2.48446	1	1.2	no result	7.19	3.595	5.295	high	6
HIST1H3G	p.P39S	4	0	0	3/5	0.5	10/10	9.50198	1	2	-0.27 (33.97%)	1.48	2.2026	4.7026	medium	7
PIGZ	p.D60E	7.5	2.146	0.5	5/6	1.5	7/7	2.16102	1	2	0.29 (71.62%)	-0.34	0.3976	4.3976	medium	8
NPBWR1	p.L252V	6	0	0.5	4/8	0.5	3/4	0.902835	0.574803	1.4	0.46 (78.46%)	1.28	1.0708	3.4708	medium	9
ORM1	p.K138N	3	0.077	0	1/8	0	3/4	0.000732283	0	0.9	0.22 (68.13%)	0.92	1.0974	1.9974	low	10
ZNF132	p.G455R	5.5	0.981	0	4/7	0.5	4/4	0.115315	0	1.1	-0.42 (25.79%)	-2.7	0.1342	1.7342	low	11

Table A.7: ***In silico*** **assessment of pathogenicity results - FALSmq2 candidate mutations.**

Gene	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**	**Priority category**	**Priority Ranking**
		Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI alignment	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score			
STRN4	p.D362E	9.5	9.816	1.5	2/8	0	2/2	-0.130677	0.905512	1.1	-1.29 (5.08%)	2.67	3.2334	5.8334	high	1
EHBP1	p.Q619L	9	5.389	1.5	4/8	0.5	1/2	3.15914	1	1.2	-0.08 (47.22%)	0.06	1.0856	4.2856	medium	2
ZFHX2	p.T565Rfs*19	6.5	0.618	0.5	1/7	2	2/2	1.71576	1	1.6	no result	no result	0	4.1	medium	3
CHRNA2	p.E411Q	6	0.079	0.5	1/8	0	3/6	0.00603937	0	0.9	-0.75 (13.67%)	0.15	1.8016	3.2016	medium	4
TUSC5	p.A142T	4	0	0	6/8	1	4/4	5.73009	1	1.6	1.26 (93.53%)	-0.82	0	2.6	medium	5
EMP2	p.I123S	7.5	3.061	0.5	1/8	0	2/6	8.74517	0.952756	1	-0.36 (28.93%)	-1.03	0.9064	2.4064	medium	6
DPH6	p.I219V	7	0.746	0.5	3/8	0	8/10	5.67361	1	1.8	no reuslt	0.05	0.025	2.325	medium	7
ALPK1	p.D979N	4.5	0.897	0	5/8	1	1/2	6.49267	1	1.2	1.46 (95.18%)	-1	0	2.2	medium	8
P2RY2	p.W16R	4.5	0.118	0	1/7	0	3/4	0.0520709	0	0.9	0.67 (84.61%)	1.29	0.9528	1.8528	low	9
SLC25A21	p.P148S	3.5	0.119	0	4/7	0.5	4/9	9.477	1	1	0.73 (86.08%)	-0.26	0.1484	1.6484	low	10
PCDHB11	p.S759T	6.5	0.103	0.5	0/8	0	2/2	-0.182402	0	0.6	1.01 (90.8%)	0.46	0.414	1.514	low	11
CFH	p.A421G	5.5	4.697	0	1/8	0	0/7	0.678087	0.0551181	0.5	0.52 (80.37%)	0.56	0.6726	1.1726	low	12
FANCC	p.D197E	4.5	0.718	0	0/8	0	0/5	1.1996	0.929134	1	0.35 (74.58%)	-1.15	0	1	low	13
ANKRD18B	p.L589R	N/A	0.122	0	0/5	0	2/2	-0.271079	0	0.6	no result	-1.5	0	0.6	low	14
CFAP47	p.Q32H	4.5	0	0	0/7	0	1/3	0.00774016	0	0.5	no reuslt	-0.15	0	0.5	low	15
CFAP47	p.D33H	4.5	0	0	3/7	0.5	1/3	-0.372622	0	0	no reuslt	-0.15	0	0.5	low	16

Table A.8: ***In silico*** **assessment of pathogenicity results - FALSmq20 candidate mutations.**

Gene	**Amino acid change**	**Gene expression**			**Protein predictons**		**Conservation**				**Genic Tolerance**			**Total score (out of 10)**	**Priority category**	**Priority Ranking**
		Brain	Spinal cord	Score	No. damaging predictions	Score	NCBI alignment	PhyloP	PhasCons	Score	RVIS	ExAC constraint Z score	Score			
RASGRF1	p.S34W	10	14.725	2	7/7	2	3/4	6.71334	1	1.4	-1.41 (4.14%)	5.31	4	9.4	high	1
NCOR2	p.R2146Q	8.5	11.999	2	6/7	2	3/3	3.41813	0.992126	1.6	-2.6 (0.82%)	2.09	3.0286	8.6286	high	2
TAZ	p.P10R	6.5	11.914	1.5	8/8	2	9/11	6.45269	1	1.8	0.08 (59.76%)	2.242	1.9258	7.2258	high	3
HIC2	p.T526M	4.5	0.598	0	6/7	2	4/4	5.96104	1	1.6	-0.4 (26.85%)	2.68	2.803	6.403	high	4
CRIM1	p.G994R	9	4.965	1	6/8	1	3/3	7.842	1	1.6	-1.1 (6.93%)	0.8	2.2614	5.8614	high	5
SLC35A4	p.C285R	9	10.109	2	7/7	2	3/5	5.65961	1	1.4	0.37 (75.29%)	-0.42	0.2842	5.6842	high	6
ELFN2	p.H303Y	no result	1.45	0	1/8	0	5/9	1.96453	0.937008	1.6	-1.84 (2.08%)	3.91	3.9134	5.5134	high	7
POU2F2	p.P82L	5.5	0.794	0	5/7	1	6/7	2.7842	0.992126	1.8	-0.54 (20.26%)	2.18	2.6848	5.4848	high	8
DNAJC4	p.Q98X	7	19.902	1.5	2/2	2	2/4	2.38284	0.952756	1.2	0.06 (58.53%)	-0.12	0.7694	5.4694	high	9
NUDC	p.Q203H	6	28.734	1.5	5/7	1	4/10	4.53506	1	1	-0.52 (21.2%)	0.62	1.886	5.386	high	10
SLC24A2	p.V286I	11.5	12.11	2	2/7	0	5/5	3.16146	1	1.8	-0.51 (21.73%)	-0.55	1.2904	5.0904	high	11
MAP1A	p.K361R	11.5	32.237	2	4/8	0.5	7/7	9.06497	1	2	1.16 (92.61%)	0.42	0.3578	4.8578	medium	12
CSMD3	p.P2472R	7.5	0.284	0.5	3/7	0.5	6/7	3.38616	1	1.8	-3.49 (0.35%)	no result	1.993	4.793	medium	13
OPRK1	p.C181S	6.5	0.04	0.5	5/7	1	3/4	2.34697	1	1.4	-0.29 (33.2%)	0.82	1.746	4.646	medium	14
SOX15	p.R119Q	5	2.373	0	7/8	2	4/4	1.75934	0.968504	1.6	no result	1.9	0.95	4.55	medium	15
COL3A1	p.L880I	5	0.836	0	3/8	0	4/4	0.719346	0.811024	1.6	-0.23 (36.34%)	3.15	2.8482	4.4482	medium	16
TSN	p.R132C	7.5	13.714	1.5	2/5	0.5	no result	-0.727189	0.0866142	0	-0.08 (47.79%)	2.71	2.3992	4.3992	medium	17
SARAF	p.K43E	11.5	137.715	2	1/7	0	2/4	3.06345	1	1.2	-0.4 (26.53%)	-0.58	1.1794	4.3794	medium	18
MIEF1	p.A36V	7	3.151	0.5	5/7	1	5/5	7.28213	1	1.8	0.42 (77.16%)	1.2	1.0568	4.3568	medium	19
TMEM199	p.A14G	8	4.177	1	5/8	1	5/6	3.86296	0.992126	1.6	0.37 (75.29%)	0.33	0.6592	4.2592	medium	20
ENPP5	p.T242M	9.5	2.63	1	5/7	1	4/4	6.17759	1	1.6	0.18 (66.07%)	-0.16	0.5986	4.1986	medium	21
TSSK4	p.A205T	4.5	0.663	0	7/8	2	4/5	3.96716	1	1.6	0.44 (77.7%)	0.3	0.596	4.196	medium	22
SLC7A14	p.E718K	8.5	4.24	1	1/8	0	3/4	5.48799	1	1.4	-1.08 (7.24%)	-0.23	1.7402	4.1402	medium	23
SLC7A14	p.E715G	8.5	4.24	1	2/8	0	3/4	4.37809	1	1.4	-1.08 (7.24%)	-0.23	1.7402	4.1402	medium	24
TTN	p.A18481T	4.5	0.127	0	6/7	2	7/7	7.71909	1	2	2.17 (98.04%)	-5.48	0	4	medium	25
LHX1	p.L324V	5	0.251	0	2/7	0	7/7	0.756756	0.204724	1.5	0.15 (64.51%)	3.54	2.4798	3.9798	medium	26
E2F8	p.H98P	4.5	0	0	3/7	0.5	6/8	3.00849	0.992126	1.8	-0.77 (13.1%)	-0.21	1.633	3.933	medium	27
ECE2	p.G194S	2	2.313	0	2/7	0	3/5	2.63561	1	1.6	-0.87 (10.59%)	0.3	1.9382	3.5382	medium	28
IRX6	p.E128G	5	0.043	0	4/8	0.5	4/5	8.47239	1	1.6	-0.75 (13.58%)	-0.7	1.3784	3.4784	medium	29
REG3G	p.I104N	4	0	0	7/8	2	3/4	1.90645	1	1.4	0.55 (81.38%)	-1.4	0	3.4	medium	30

ALDH3B1	unknown	5	3.171	0	1/1	2	no result	6.99356	1	1	no result	0.59	0.295	3.295	medium	31
DNAJC13	p.K1277E	8.5	3.601	1	3/8	0	2/6	9.32576	1	1	0.17 (65.77%)	1.04	1.2046	3.2046	medium	32
MARVELD2	p.E454G	no result	0.338	0	6/8	1	2/4	6.63418	1	1.2	-0.35 (29.43%)	-1.07	0.8764	3.0764	medium	33
BAHCC1	unknown	7	no result	0.5	5/6	1.5	no result	2.36606	0.992126	1	no result	no result	0	3	medium	34
USP53	p.S606N	6	4.215	0.5	4/8	0.5	6/6	5.93234	1	1.8	0.83 (88.11%)	-0.3	0.0878	2.8878	medium	35
RDH12	p.P38S	5.5	0.118	0	1/8	0	2/9	1.12308	1	1	-0.65 (16.44%)	0.36	1.8512	2.8512	medium	36
OR4Q3	p.I45T	no result	0	0	5/8	1	3/3	3.67212	0.84252	1.6	0.8 (87.59%)	no result	0.2482	2.8482	medium	37
OR2K2	p.S238F	4.5	0.395	0	6/8	1	6/6	2.70683	0.984252	1.8	0.33 (73.61%)	-1.25	0	2.8	medium	38
RSRP1	p.A270G	7.5	19.82	1.5	0/8	0	no result	-0.199386	0	0	0.91 (89.44%)	2.17	1.2962	2.7962	medium	39
ABHD15	p.A41T	5.5	1.508	0	1/7	0	1/3	0.033685	0	0.5	-0.25 (36.07%)	1.86	2.2086	2.7086	medium	40
FCHSD1	p.R671H	5.5	3.365	0	4/8	0.5	3/3	3.78759	0.992126	1.6	0.23 (68.54%)	-0.16	0.5492	2.6492	medium	41
OR4D9	p.S74P	3.5	0	0	5/8	1	3/3	0.219362	0	1.1	1.66 (96.28%)	-1.33	0	2.1	medium	42
HR	p.R582Q	7.5	3.681	0.5	1/8	0	3/4	0.997157	0.976378	1.4	1.24 (93.31%)	-0.12	0.0738	1.9738	low	43
KIF26A	p.G61S	4.5	0.411	0	1/8	0	4/5	0.367575	0.661417	1.6	no result	-1.21	0	1.6	low	44
PNMAL2	p.T226S	no result	5.203	0.5	3/7	0.5	5/6	-0.0650551	0.0157323	0.6	no result	no result	0	1.6	low	45
LMTK3	p.P641L	no result	2.587	0	3/7	0.5	4/5	2.2099	0.244094	1.1	no result	no result	0	1.6	low	46
FCGBP	p.P3983L	5	0.594	0	5/8	1	1/8	0.608126	0	0.5	no result	no result	0	1.5	low	47
MGAM	p.W797R	4	0.07	0	5/7	1	1/4	0.334512	0.275591	0.5	2.36 (98.45%)	-2.46	0	1.5	low	48
MRPS28	p.Q64P	5.5	8.613	0.5	1/7	0	2/4	-0.0563071	0.850394	0.7	0.9 (89.39%)	-0.21	0.1072	1.3072	low	49
ERVV-1	p.Y390F	no result	0	0	1/3	0	2/2	0.326157	0.497386	1.1	no result	no result	0	1.1	low	50
MROH5	unknown	5	0	0	0/3	0	no result	2.04943	0.724409	1	4.38 (99.74%)	no result	0.0052	1.0052	low	51
SLC22A24	p.C385Y	4	0	0	0/7	0	0/3	1.13282	0.992126	1	no result	-1.84	0	1	low	52
MXRA5	p.D2583N	4.5	0.129	0	2/8	0	2/6	3.14629	0.992126	1	1.57 (95.68%)	-1.92	0	1	low	53
CGREF1	p.V313M	7	1.269	0.5	2/8	0	0/3	-0.291882	0.00787402	0	0.44 (77.8%)	-0.11	0.389	0.889	low	54
FASTKD2	p.R153H	7	2.961	0.5	2/8	0	0/3	-0.736362	0	0	0.87 (88.8%)	0	0.224	0.724	low	55
PLEKHG4B	p.A539S	5	0	0	0/8	0	2/3	-0.774732	0.00787402	0.2	-0.1 (45.61%)	-1.22	0.4778	0.6778	low	56
FANCA	p.S1301P	4.5	0.231	0	0/8	0	4/5	-1.6082	0	0.6	-0.1 (45.67%)	-5.81	0	0.6	low	57
LOC79999	p.Q64H	no result	no result	0	1/2	0.5	no result	-0.0487402	0	0	no result	no result	0	0.5	low	58
DNAH11	p.P1792L	4	0	0	1/8	0	0/4	0.999811	0	0.5	no result	no result	0	0.5	low	59
PLEKHG4B	p.V504M	5	0	0	0/8	0	1/3	-0.329898	0	0	-0.1 (45.61%)	-1.22	0.4778	0.4778	low	60
LIPF	p.G27R	4	0.021	0	0/8	0	2/6	-1.20017	0	0	0.51 (80.01%)	-0.24	0.2798	0.2798	low	61
MUC16	p.R2358Q	no result	0	0	0/7	0	1/2	-2.02506	0	0.2	29.75 (100%)	no result	0	0.2	low	62
CEP295	p.K1491E	no result	1.349	0	0/8	0	2/5	-0.562929	0	0	5.12 (99.83%)	-0.17	0	0	low	63
KRTAP29-1	p.M137K	no result	0	0	0/6	0	1/3	-0.544693	0	0	no result	no result	0	0	low	64

A.3.5 Supportive *in silico* data collected for family candidate mutations

The following tables contain supporrtive *in silico* data collected for the candidate mutations (and the gene in which they reside) identified in families FALS15, FALS45, FALSmq2 and FALSmq20.

Table A.9: **Data to support the potential pathogenicity of each candidate mutation from FALS15.**

Gene	Amino acid change	Gene name	GeneCards description	PubMed matches with neurodegenerative disease	ALS linked protein interacting partners	SMART domain	Changes to NetPhos2.0 phosphorylation sites
CLCN4	p.I668T	Chloride Voltage-Gated Channel 4	Voltage-dependent chloride channel genes	none	none	unknown region	no change
MTSS1L	p.A126T	Metastasis Suppressor 1-Like	Associated with actin binding and cytoskeletal adaptor activity	none	FUS	Coiled coil domain	no change
SCN4A	p.R225W	Sodium Voltage-Gated Channel Alpha Subunit 4	Member of the sodium channel alpha subunit gene family. Responsible for the generation and propagation of action potentials in neurons and muscle	none	none	Transmembrane helix region	no change
LRRN2	p.I196T	Leucine Rich Repeat Neuronal 2	Leucine-rich repeat protein, showing homology with cell-adhesion molecules or as signal transduction receptors	1	none	Leucine rich repeat domain	no change
SUPV3L1	p.Q168E	Suv3 Like RNA Helicase	Associated with RNA binding and RNA binding	none	HNRNPA1	unknwon region	no change
HOXD3	p.Y249C	Homeobox D3	Homeobox protein; conserved transcription factor	none	none	HOX Homeodomain	no change
FAM171A1	p.H518R	Family With Sequence Similarity 171 Member A1	None available	none	none	unknwon region	no change
SP1	p.A145T	Sp1 Transcription Factor	Zinc finger transcription factor	41	PURA, SFPQ	unknwon region	no change
MAPKAPK3	p.K368R	MAPK-Activated Protein Kinase 3	Mitogen-activated protein (MAP) kinase	none	none	unknwon region	S373 added
SIM1	p.G733V	Single-Minded Family BHLH Transcription Factor 1	Potentially involved in abnormal developmental processes	3	no result	unknwon region	no change
ZNF385B	p.P435S	Zinc Finger Protein 385B	Associated with nucleic acid binding and p53 binding	none	none	low complexity	no change
TYMP	p.Q245E	Thymidine Phosphorylase	Promotes angiogenesis	none	none	unknwon region	no change
TNS2	p.S992L	Tensin 2	Tensin protein that binds to actin filaments and participates in signaling pathways. Regulates cell migration	none	none	low complexity	S992 and S994 removed
NECAB3	p.R203L	N-Terminal EF-Hand Calcium Binding Protein 3	May regulate amyloid precursor protein metabolism and beta-amyloid generation	none	none	low complexity	S205 removed
ZNF425	p.R424P	Zinc Finger Protein 425	Associted with nucleic acid binding	none	none	Zinc finger	no change
CEP295	p.N1707I	Centrosomal Protein 295	Mediates centriole-to-centrosome conversion during late mitosis	none	none	unknwon region	no change
ZNF497	p.K23R	Zinc Finger Protein 497	Potentially involved in transcriptional regulation	none	none	unknwon region	no change
ZNF497	p.V22G	Zinc Finger Protein 497	Potentially involved in transcriptional regulation	none	none	unknwon region	no change
RNF133	p.R94Q	Ring Finger Protein 133	RING finger protein	none	no result	unknown region	no change

Table A.10: **Data to support the potential pathogenicity of each candidate mutation from FALS45.**

Gene	Amino acid change	Gene name	GeneCards description	PubMed matches with neurodegenerative disease	ALS linked protein interacting partners	SMART domain	Changes to NetPhos2.0 phosphorylation sites
SCCPDH	p.V256L	Saccharopine Dehydrogenase (Putative)	Associated with oxidoreductase activity	none	SQSTM1, UBC	unknwon	S254 removed
GDPD1	p.P221T	Glycerophosphodiester Phosphodiesterase Domain Containing 1	Catalyses the hydrolysis of deacylated glycerophospholipids to glycerol	none	UBC	transmembrane region	T221 added
SPATA2	p.G206S	Spermatogenesis Associated 2	None available	none	none	unknwon	S206 added
KRT85	p.S5P	Keratin 85	Keratin protein	none	UBC	unknwon	no change
GABRG3	p.S236F	Gamma-Aminobutyric Acid Type A Receptor Gamma3 Subunit	Gamma subunit of gamma-aminobutyric acid (GABA) receptor; the major inhibitory neurotransmitter in the brain	1	none	unknwon	S236 and T235 removed
GRIN2D	p.V144L	Glutamate Ionotropic Receptor NMDA Type Subunit 2D	Subunit of the N-methyl-D-aspartate (NMDA) receptor	2	VCP	unknwon	no change
HIST1H3G	p.P39S	Histone Cluster 1 H3 Family Member G	Replication-dependent histone which belongs to the H3 family of histones	none	none	Histone H3	S87 added
PIGZ	p.D60E	Phosphatidylinositol Glycan Anchor Biosynthesis Class Z	Involved in glycosylphosphatidylinositol (GPI) anchor biosynthesis	none	none	unknwon	no change
NPBWR1	p.L252V	Neuropeptides B/W Receptor 1	Associated with Peptide ligand-binding receptors and Signaling by GPCR	none	none	transmembrane region	T250 added
ORM1	p.K138N	Orosomucoid 1	Plasma protein which increases in response to inflammation	1	none	unknwon	S143 removed
ZNF132	p.G455R	Zinc Finger Protein 132	Associated with nucleic acid binding and transcription factor activity, sequence-specific DNA binding	none	none	unknwon	no change

Table A.11: **Data to support the potential pathogenicity of each candidate mutation from FALSmq2.**

Gene	Amino acid change	Gene name	GeneCards description	PubMed matches with neurodegenerative disease	ALS linked protein interacting partners	SMART domain	Changes to NetPhos2.0 phosphorylation sites
STRN4	p.D362E	Striatin 4	Involved in calmodulin binding	none	DCTN1, EEF1A1, NONO	unknown region	No change
EHBP1	p.Q619L	EH Domain Binding Protein 1	Eps15 homology domain binding proteinwith a potential role in endocytic trafficking	none	none	unknown region	No change
ZFHX2	p.T565Rfs*19	Zinc Finger Homeobox 2	Zinc Finger homeobox protein associated with nucleic acid binding and actin binding	none	none	uknown region	S565 added
CHRNA2	p.E411Q	Cholinergic Receptor Nicotinic Alpha 2 Subunit	Alpha subunit of a cotinic acetylcholine receptor (muscle and neuronal receptor)	4	none	low complexity region	S406 removed
TUSC5	p.A142T	Tumor Suppressor Candidate 5	Associated with Accommodative Esotropia and Chiasmal Syndrome	none	none	unknown region	T142 added
EMP2	p.I123S	Epithelial Membrane Protein 2)	Tetraspan protein which regulates cell membrane composition	4	none	unknown region	S138 added
DPH6	p.I219V	Diphthamine Biosynthesis 6	Involved in transport to the Golgi Apparatus	1	none	unknown region	No change
ALPK1	p.D979N	Alpha Kinase 1	An alpha kinase protein	none	none	unknown region	No change
P2RY2	p.W16R	Purinergic Purinergic Receptor P2Y2	A P2 receptor involved in proliferation, apoptosis and inflammation	5	none	unknown region	No change
SLC25A21	p.P148S	Solute Carrier Family 25 Member 21	Mitochondrial carrier transporting oxodicarboxylates	1	FBXO6	unknown region	No change
PCDHB11	p.S759T	Protocadherin Beta 11	Neural cadherin-like cell adhesion protein, integral to the plasma membrane	none	none	unknown region	No change
CFH	p.A421G	Complement Factor H	Involved in the regulation of complement activation	23	none	CCP (complement control protein)	No change
FANCC	p.D197E	FA Complementation Group C	Member of the Fanconi anemia complementation group C	1	CDK1	unknown region	No change
ANKRD18B	p.L589R	Ankyrin Repeat Domain 18B	Associated with nucleotide binding	none	none	Coiled coil region	S1277 removed
CFAP47	p.Q32H	Cilia And Flagella Associated Protein 47	None available	none	none	unknown region	Y31 removed
CFAP47	p.D33H	Cilia And Flagella Associated Protein 47	None available	none	none	unknown region	Y31 removed

Table A.12: **Data to support the potential pathogenicity of each candidate mutation from FALSmq20.**

Gene	Amino acid change	Gene name	GeneCards description	PubMed matches with neurodegenerative disease	ALS linked protein interacting partners	SMART domain	Changes to NetPhos2.0 phosphorylation sites
RASGRF1	p.S34W	Ras Protein Specific Guanine Nucleotide Releasing Factor 1	Guanine nucleotide exchange factor; stimulates the dissociation of GDP from RAS protein	9	none	Pleckstrin homology domain	no change
NCOR2	p.R2146Q	Nuclear Receptor Corepressor 2	Mediates transcriptional silencing of certain target genes	3	none	unknown region	no change
TAZ	p.P10R	Tafazzin	Highly expressed in cardiac and skeletal muscle; associated with various cardiac related diseases	4	none	unknown region	no change
HIC2	p.T526M	Hypermethylated In Cancer 2	Associated with C-terminus binding	1	none	Zinc finger domain	no change
CRIM1	p.G994R	Cysteine Rich Transmembrane BMP Regulator 1	Transmembrane protein; may play a role in tissue development	none	none	unknown region	no change
SLC35A4	p.C285R	Solute Carrier Family 35 Member A4	Associated with sugar:proton symporter activity	none	none	Transmembrane region	no change
ELFN2	p.H303Y	Extracellular Leucine Rich Repeat And Fibronectin Type III Domain Containing 2)	Associated with phosphatase binding and protein phosphatase inhibitor activity	none	UBC	Fibronectin type 3 domain	Y303 added
POU2F2	p.P82L	POU Class 2 Homeobox 2	Homeobox-containing transcription factor of the POU domain family	none	none	low complexity region	no change
DNAJC4	p.Q98X	DnaJ Heat Shock Protein Family (Hsp40) Member C4	Associated with unfolded protein binding	none	none	unknown region	no sequnce result
NUDC	p.Q203H	Nuclear Distribution C, Dynein Complex Regulator	Involved in spindle formation during mitosis and in microtubule organisation during cytokinesis	1	SOD1 and VPS29	unknown region	no change
SLC24A2	p.V286I	Solute Carrier Family 24 Member 2	Transporter protein belonging to the calcium/cation antiporter superfamily	none	none	unknown region	no change
MAP1A	p.K361R	Microtubule Associated Protein 1A	Thought to be involved in microtubule assembly; expression almost exclusively in the brain	15	none	Coiled coil domain	no change
CSMD3	p.P2472R	CUB And Sushi Multiple Domains 3	Associated with Benign Adult Familial Myoclonic Epilepsy and Trichorhinophalangeal Syndrome	none	UBC	Complement control protein module	Y2475 removed
OPRK1	p.C181S	Opioid Receptor Kappa 1	Opioid receptor	1	none	Transmembrane region	no change
SOX15	p.R119Q	SRY-Box 15	Member of the SOX (SRY-related HMG-box) family of transcription factors involved in the regulation of embryonic development and in the determination of the cell fate	none	none	unknown region	no change
COL3A1	p.L880I	Collagen Type III Alpha 1 Chain	A fibrillar collagen found in extensible connective tissues	1	none	low complexity region	no change
TSN	p.R132C	Translin	DNA-binding protein involved in chromosomal translocations	3	VPS29	unknown region	no change
SARAF	p.K43E	Store-Operated Calcium Entry Associated Regulatory Factor	None available	none	UBC	unknown region	no change
MIEF1	p.A36V	Mitochondrial Elongation Factor 1	associated with identical protein binding and ADP binding	1	UBQLN1	Transmembrane region	no change
TMEM199	p.A14G	Transmembrane Protein 199	Localise to the endoplasmic reticulum (ER)-Golgi intermediate compartment and coat protein complex I	none	none	unknown region	no change

ENPP5	p.T242M	Ectonucleotide Pyrophosphatase/Phosphodiesterase 5 (Putative)	Type-I transmembrane glycoprotein; may play a role in neuronal cell communications	none	none	unknown region	T242 removed
TSSK4	p.A205T	Testis Specific Serine Kinase 4	Member of the testis-specific serine/threonine kinase family, may be involved in involved in spermatogenesis	none	none	Catalytic domain	no change
SLC7A14	p.E718K	Solute Carrier Family 7 Member 14	Primarily expressed in skin, neural tissue, and primary endothelial cells; predicted to mediate lysosomal uptake of cationic amino acids	1	none	unknown region	no change
SLC7A14	p.E715G	Solute Carrier Family 7 Member 14	Primarily expressed in skin, neural tissue, and primary endothelial cells; predicted to mediate lysosomal uptake of cationic amino acids	1	none	unknown region	no change
TTN	p.A18481T	Titin	A large protein; abundant striated muscle	9	SQSTM1	Fibronectin type 3 domain	T24978 added
LHX1	p.L324V	LIM Homeobox 1	Transcription factor important for the development of the renal and urogenital systems	none	none	low complexity region	S320 removed; T322 added
E2F8	p.H98P	E2F Transcription Factor 8	Regulates gene expression during the cell cycle	none	EWSR1	unknown region	no change
ECE2	p.G194S	Endothelin Converting Enzyme 2	Membrane-bound zinc-dependent metalloprotease	1none	UBC	Transmembrane region	no change
IRX6	p.E128G	Iroquois Homeobox 6	Associated with sequence-specific DNA binding	none	none	unknown region	no change
REG3G	p.I104N	Regenerating Family Member 3 Gamma	Antimicrobial lectin protein	1	none	C-type lectin /carbohydrate-recognition domain	S100 added
ALDH3B1	unknown	Aldehyde Dehydrogenase 3 Family Member B1	Oxidises long-chain fatty aldehydes; may play a role in protection from oxidative stress	none	none	no result	no sequnce result
DNAJC13	p.K1277E	DnaJ Heat Shock Protein Family (Hsp40) Member C13	Plays a role in clathrin-mediated endocytosis,may also be involved in post-endocytic transport mechanisms	1none	none	unknown region	no change
MARVELD2	p.E454G	MARVEL Domain Containing 2	Helps establish epithelial barriers	none	none	unknown region	Y458 removed
BAHCC1	unknown	BAH Domain And Coiled-Coil Containing 1	Associated with chromatin binding	none	none	no result	no sequnce result
USP53	p.S606N	Ubiquitin Specific Peptidase 53	Associated with thiol-dependent ubiquitinyl hydrolase activity	1	none	low complexity region	no change
RDH12	p.P38S	Retinol Dehydrogenase 12	NADPH-dependent retinal reductase; involved in the metabolism of short-chain aldehydes	12	UBC	unknown region	S38 added
OR4Q3	p.I45T	Olfactory Receptor Family 4 Subfamily Q Member 3	Olfactory receptor; involved in the neuronal response that triggers the perception of smell	none	none	Transmembrane region	no change
OR2K2	p.S238F	Olfactory Receptor Family 2 Subfamily K Member 2	Olfactory receptor; involved in the neuronal response that triggers the perception of smell	none	none	Transmembrane region	S238 removed
RSRP1	p.A270G	Arginine And Serine Rich Protein 1	None available	none	none	low complexity region	no change
ABHD15	p.A41T	Abhydrolase Domain Containing 15	Associated with hydrolase activity	none	none	low complexity region	T41 added
FCHSD1	p.R671H	FCH And Double SH3 Domains 1	None available	none	none	low complexity region	no change
OR4D9	p.S74P	Olfactory Receptor Family 4 Subfamily D Member 9	Olfactory receptor; involved in the neuronal response that triggers the perception of smell	none	none	unknown region	no change

HR	p.R582Q	HR, Lysine Demethylase And Nuclear Receptor Corepressor	Transcriptional corepressor of multiple nuclear receptors involved in hair growth	3	none	low complexity region	no change
KIF26A	p.G61S	Kinesin Family Member 26A	Involved in Platelet activation, signaling and aggregation and Vesicle-mediated transport	none	none	unknown region	no change
PNMAL2	p.T226S	Paraneoplastic Ma Antigen Family Like 2	None available	none	none	unknown region	no change
LMTK3	p.P641L	Lemur Tyrosine Kinase 3	Assocoaited with transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity	none	none	low complexity region	no change
FCGBP	p.P3983L	Fc Fragment Of IgG Binding Protein	None available	none	none	unknown region	no change
MGAM	p.W797R	Maltase-Glucoamylase	Maltase-glucoamylase which plays a role in the final steps of digestion of starch	none	none	unknown region	no change
MRPS28	p.Q64P	Mitochondrial Ribosomal Protein S28	Mitochondrial ribosomal protein	none	none	unknown region	no change
ERVV-1	p.Y390F	Endogenous Retrovirus Group V Member 1, Envelope	Part of a human endogenous retrovirus (HERV) family of proteins; important in reproduction	none	none	unknown region	no change
MROH5	unknown	Maestro Heat Like Repeat Family Member 5	Associated with binding	none	none	no result	no sequnce result
SLC22A24	p.C385Y	Solute Carrier Family 22 Member 24	Transmembrane protein; involved in transport organic ions across cell membranes	none	EWSR1	Transmembrane region	no change
MXRA5	p.D2583N	Matrix Remodeling Associated 5	Matrix-remodelling associated protein	none	none	Immunoglobulin C-2 type domain	no change
CGREF1	p.V313M	Cell Growth Regulator With EF-Hand Domain 1	Associated with calcium ion binding	none	none	unknown region	no change
FASTKD2	p.R153H	FAST Kinase Domains 2	May play a role in mitochondrial apoptosis	3	HNRNPA1	unknown region	S155 removed
PLEKHG4B	p.A539S	Pleckstrin Homology And RhoGEF Domain Containing G4B	Associated with Rho guanyl-nucleotide exchange factor activity	none	none	unknown region	no change
FANCA	p.S1301P	FA Complementation Group A	Member of the Fanconi anemia complementation group A	1	CDK1	unknown region	S1301 removed
LOC79999	p.Q64H	Uncharacterised LOC79999	None available	none	none	no result	no sequnce result
DNAH11	p.P1792L	Dynein Axonemal Heavy Chain 11	Microtubule-dependent motor ATPase; reportedly involved in the movement of respiratory cilia	none	none	low complexity region	no change
PLEKHG4B	p.V504M	Pleckstrin Homology And RhoGEF Domain Containing G4B	Associated with Rho guanyl-nucleotide exchange factor activity	none	none	unknown region	no change
LIPF	p.G27R	Lipase F, Gastric Type	Hydrolyses the ester bonds of triglycerides as part of digestion of dietary triglycerides	none	none	unknown region	no change
MUC16	p.R2358Q	Mucin 16, Cell Surface Associated	Associated with Ovarian Cancer and Childhood Ovarian Cancer	none	UBC	unknown region	T2360 removed
CEP295	p.K1491E	Centrosomal Protein 295	None available	none	none	unknown region	no change
KRTAP29-1	p.M137K	Keratin Associated Protein 29-1	None available	none	none	unknown region	no change

A.4 Additional figures

#CHROM	POS	POS_TO
chr1	225044284	225944165
chr1	238783484	239931618
chr3	158030382	160325458
chr3	177177991	179005976
chr4	12939091	13526386

Figure A.1: **Regions file example.**Example of the tab-delimited file used to define the genomic regions to be subset using BCFTools.

2	41	0	0	1	1	0	0	0	0	0	0	0	0
2	42	0	0	2	2	0	0	0	0	0	0	0	0
2	43	0	0	1	1	0	0	0	0	0	0	0	0
2	44	41	42	2	2	0	0	0	0	0	0	0	0
2	45	41	42	1	2	0	0	0	0	0	0	0	0
2	46	0	0	2	1	0	0	0	0	0	0	0	0
2	47	43	44	1	2	0	0	0	0	0	0	0	0
2	48	43	44	1	0	3	3	1	2	2	2	2	2
2	49	43	44	2	0	3	3	1	2	2	2	2	2
2	50	43	44	2	0	3	3	1	2	4	2	2	2
2	51	0	0	2	1	0	0	0	0	0	0	0	0
2	52	47	51	2	0	3	3	1	2	4	2	2	2
2	53	47	51	2	0	3	3	1	1	2	2	2	2
2	54	47	51	1	2	3	3	1	1	2	2	2	2
2	55	0	0	2	1	3	3	2	2	4	2	2	2
2	56	54	55	1	0	3	3	1	2	2	2	2	2
2	57	54	55	1	0	3	3	1	2	2	2	2	2
2	58	45	46	2	2	3	3	1	1	2	2	2	2
2	59	0	0	1	1	0	0	0	0	0	0	0	0
2	60	59	58	1	2	1	3	1	1	4	2	2	2
2	61	0	0	2	1	3	3	1	2	2	2	2	2
2	62	59	58	2	0	3	3	1	1	2	2	2	2
2	63	59	58	2	0	1	3	1	1	4	2	2	2
2	64	59	58	2	0	3	3	1	1	2	2	2	2
2	65	59	58	2	0	1	3	1	1	4	2	2	2

Figure A.2: **Example of a ped file used for linkage analysis using Merlin software.** Each line represents an individual member of the family (pedigree). The first six columns define relevant details about the individual. From one to six, each column defines the individual's family membership, individual ID, mother's ID, father's ID, sex (1=male, 2=female) and affected status (1=unaffected control, 2=affected, 0=at-risk/unknown). The final column indicates the liability class of each individual (according to Table 6.3). The intervening columns each represent the identity of one allele of a SNP marker (1=A, 2=C, 3=G, 4=T, 0=unknown), in pairs.

```
A       Affected.status
M       exm5488
M       rs2843160
M       rs3736330
M       rs2494427
M       rs7515934
M       rs1107685
M       rs2494625
M       exm6141
M       rs10909793
```

FIGURE A.3: **Example of a dat file used for linkage analysis using Merlin software.** Each line defines the identity of a column in the corresponding ped file, following the first five basic (invariable) ped file columns. Column one indicates the data type (A=affection status, M=marker, C=Covariate), while column two acts as a label for the column.

```
1       exm5488          0.0161264
1       rs2843160        0.0748622
1       rs3736330         0.130664
1       rs2494427          0.16631
1       rs7515934         0.232589
1       rs1107685          0.47837
1       rs2494625         0.526496
1       exm6141           0.649374
1       rs10909793         1.58045
1       rs10797342          1.5907
1       rs2842910          1.65138
1       rs2485944          1.85742
```

FIGURE A.4: **Example of a map file used for linkage analysis using Merlin software.** Each line reprsents a marker, and the columns one to three indicate chromosome, marker name (corresponding to the label in the dat file) and position (genetic distance).

```
Affected.status 0.0001  *   Dominant_Model
   LIABILITY = 1                    0.0000,0.0100,0.0100
   LIABILITY = 2                    0.0000,0.3000,0.3000
   LIABILITY = 3                    0.0000,0.500,0.500
   LIABILITY = 4                    0.0000,0.7000,0.7000
   LIABILITY = 5                    0.0000,0.85000,0.85000
   LIABILITY = 6                    0.0000,0.9000,0.9000
   OTHERWISE                   0.0000,0.0000,0.0000
```

Figure A.5: **File used to specify the disease model for parametric linkage analysis using Merlin.** The first line defines the model, which is based on unknwon affection status and a disease allele frequency of 0.0001. The following lines define the liability classes for the likelihood of an individual carrying 0, 1 or 2 disease alleles.

A.5 Co-authored publications presented in this thesis

The following publications were co-authored by the candidate. Papers A1-A3 are presented in Chapter 4, and Paper A4 is presented in Chapter 7.

A.5.1 Paper A1

A.5.2 Paper A2

A.5.3 Paper A3

A.5.4 Paper A4

van Rheenen, W., Shatunov, A., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48, p. 1043-1048.

DOI: 10.1038/ng.3622

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-00471-1 OPEN

Cross-ethnic meta-analysis identifies association of the *GPX3-TNIP1* locus with amyotrophic lateral sclerosis

Beben Benyamin *et al.*#

Cross-ethnic genetic studies can leverage power from differences in disease epidemiology and population-specific genetic architecture. In particular, the differences in linkage disequilibrium and allele frequency patterns across ethnic groups may increase gene-mapping resolution. Here we use cross-ethnic genetic data in sporadic amyotrophic lateral sclerosis (ALS), an adult-onset, rapidly progressing neurodegenerative disease. We report analyses of novel genome-wide association study data of 1,234 ALS cases and 2,850 controls. We find a significant association of rs10463311 spanning *GPX3-TNIP1* with ALS ($p = 1.3 \times 10^{-8}$), with replication support from two independent Australian samples (combined 576 cases and 683 controls, $p = 1.7 \times 10^{-3}$). Both *GPX3* and *TNIP1* interact with other known ALS genes (*SOD1* and *OPTN*, respectively). In addition, *GGNBP2* was identified using gene-based analysis and summary statistics-based Mendelian randomization analysis, although further replication is needed to confirm this result. Our results increase our understanding of genetic aetiology of ALS.

Correspondence and requests for materials should be addressed to N.R.W. (email: naomi.wray@uq.edu.au)
#A full list of authors and their affliations appears at the end of the paper

For people of European ancestry, the lifetime risk of amyotrophic lateral sclerosis (ALS) is 0.3–0.5%[1, 2], with peak age of onset of 58–63 years[3], and median survival of 2–4 years[4]. Investigations of families with multiple affected individuals have led to the identification of mutations that segregate with disease in a number of genes, including *SOD1*, *C9orf72*, *TARDBP*, *FUS* and *TBK1*[5, 6]. However, about 90% of cases[5] ('sporadic ALS' (sALS)) present with sparse or no family history. Nonetheless, genome-wide association studies (GWAS) have provided direct evidence of a genetic contribution to sALS, with estimates that ~8.5%[7] of variance in liability is tagged by common single-nucleotide polymorphisms (SNPs). Currently, only a small proportion of this variation (0.2% of variance in liability)[7] is accounted for by the six common loci (*C9orf72*, *UNC13A*, *SARM1*, *MOBP*, *SCFD1*, *C21orf2*) identified as significant based on association analysis of 12,577 cases and 23,475 controls[7]. The SNP-heritability estimate implies that more risk loci will be detected with increasing sample size, as found for other complex genetic diseases[8]. Whole-exome sequencing (WES) studies, designed to identify genes enriched for rare variants, have also been conducted for sALS. The largest study, comprising 2,874 cases and 6,405 controls, identified *TBK1* as a novel ALS risk gene[6], with GWAS support for association of common loci ($p = 6.6 \times 10^{-8}$)[7]. Rare variant burden analysis in a WES of 1,022 index familial cases identified p.Arg261His in *NEK1* as an ALS associated variant, and follow-up in large samples suggest that this variant together with *NEK1* loss of function mutations account for ~3% of ALS cases[9].

To date, the largest genetic studies for ALS are in the subjects of European ancestry, but common variants associated with disease are likely to be ancient and shared across ethnicities. Given sufficient power, cross-ethnic genetic studies can aid fine mapping of disease loci, exploiting differences in allele frequency and linkage disequilibrium (LD). In China, the lifetime risk of ALS is estimated to be lower (0.1%)[1] and its mean age of onset is estimated to be a few years earlier than in Europe[4, 10]. High penetrance mutations in known ALS genes identified in Europeans have been detected in Chinese cases[11], but the frequency of the *C9orf72* expansion is much lower (0.3%)[12] than in Europeans (frequency 7%)[5], and it may have arisen on a different haplotype background[12].

In a cross-ethnic meta-analysis of the largest GWAS for ALS in Europeans[7], together with a new Chinese data set, we identify the *GPX3-TNIP1* locus to be significantly associated with ALS ($p = 1.3 \times 10^{-8}$). This association is replicated in two independent Australian cohorts with a combined *p*-value of 1.7×10^{-3}. Previous studies indicate functional relevance of both *GPX3* and *TNIP1*[13–18]. The identification of this locus contributes to a better understanding of the genetic aetiology of ALS.

Results

Genome-wide association analysis. We conduct a genome-wide (GW) association analysis in a Chinese sample of 1,234 sALS cases and 2,850 controls (Supplementary Table 1 and Supplementary Figs 1–3). The genomic inflation factor λ_{GC} of 1.02 and λ_{1000} of 1.01 showed no evidence for inflation in test statistics. The combined effects of all common genetic variants on ALS liability (SNP-heritability) estimated from the Chinese GWAS data is 15.1% (SE): 4%; $p = 9.5 \times 10^{-5}$) using GCTA-GREML[19] and 15.0% (SE: 3.5%) using LD score regression[20] (intercept 1.0, which also shows no evidence of population stratification). Given the SE, these estimates are not different from the estimate of 8.5% (SE 0.5%) from European data[7]. Partitioning of the SNP-heritability by chromosome showed a significant positive correlation with chromosome length (Supplementary Fig. 4a) consistent with a polygenic architecture. Based on minor allele frequency (MAF) bin, the SNP-heritability was attributed to SNPs across the MAF range, but SEs per bin were large (Supplementary Fig. 4b); similar analyses based on European data suggested that less common SNPs tagged more variation than other MAF classes[7].

No individual SNPs passed the GW significant *p* value threshold of 5×10^{-8}, and none of the significant SNPs reported in the European[7] GWAS replicated in our samples ($p > 0.05$). We also checked for the associations of two GW significant SNPs in previous GWAS of Chinese cohort of ALS patients[21]. However, we could not replicate the association in that study. We note that despite evidence for population stratification, principal components derived from SNP data of the previous study were not included as covariates in their association analysis. The *p* values for rs6703183 and rs8141797 are 0.07 and 0.12 in our Chinese samples and 0.66 and 0.94 in European GWAS results, respectively. Direction of effect sign tests (Supplementary Table 2) and polygenic risk scoring analyses (Supplementary Fig. 5) provided no conclusive evidence of shared risk loci (Nagelkerke's $R^2 = 0.002$; $p = 0.01$). These results are not unexpected given the size of our sample and effect sizes estimated in Europeans. The Chinese GWAS sample had 80% power to identify common genetic variants of genotype relative risk of 1.4 and 1.8 for risk allele frequency of 0.2 and 0.05, respectively, at the GW threshold of significance $p = 5 \times 10^{-8}$.

Meta-analysis. Meta-analysis of our results with those of the European[7] GWAS identified a new GW significant locus at chromosome 5p33.1 (rs10463311, risk allele C, odds ratio (OR) 1.11 95% confidence interval (CI): 1.06–1.14, $p_{logistic} = 2.9 \times 10^{-8}$; $p_{lmm} = 1.3 \times 10^{-8}$) spanning the genes *GPX3* and *TNIP1* (Figs. 1 and 2; Table 1; Supplementary Data 1) for which the risk allele is

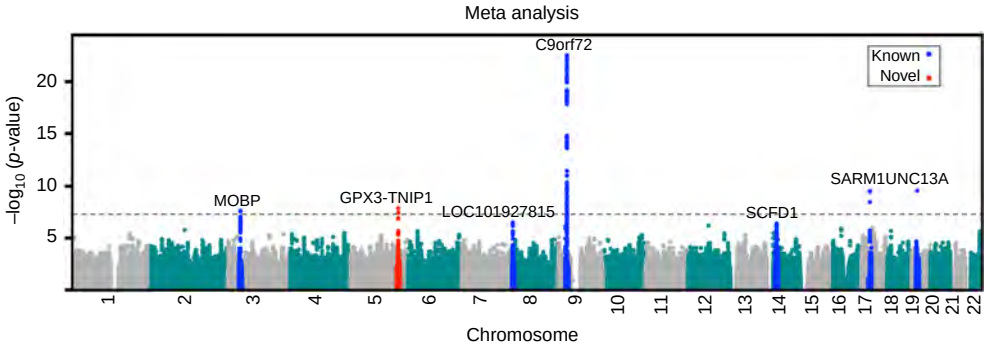


Fig. 1 Manhattan plot of the meta-analysis between European and Chinese GWAS revealed a novel locus, *GPX3-TNIP1* (*red*). Loci previously identified in the largest European GWAS are presented in *blue*. The *p* values are from the linear mixed model

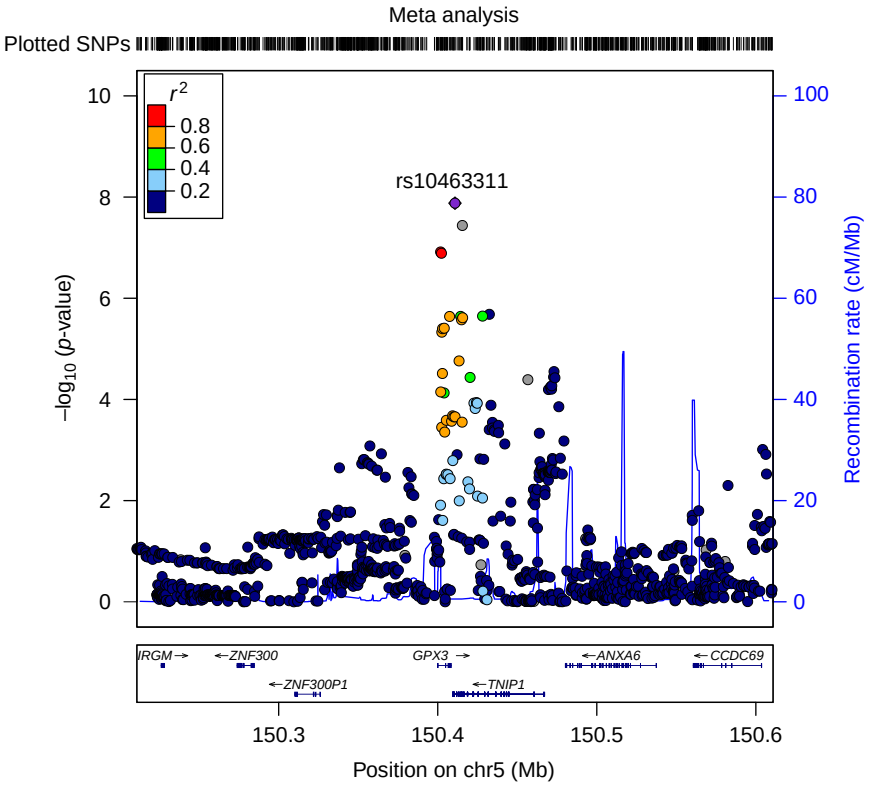


Fig. 2 Regional ALS association plot of the *GPX3-TNIP1* locus from the meta-analysis results created using LocusZoom[44]. From the meta-analysis, rs10463311 is the SNP with the strongest association with ALS ($p = 1.3 \times 10^{-8}$). This SNP is replicated in two independent Australian cohorts with combined $p = 1.7 \times 10^{-3}$

Table 1 Association analysis results between rs10463311 spanning GPX3-TNIP1 and ALS across cohorts

Cohort	*N* cases	*N* cont	Freq cases	Freq cont	OR	95% CI	$p_{logistic}$
European[2]	12,577	23,475	0.27	0.24	1.11	1.07–1.15	8.5×10^{-7}
Chinese	1,234	2,850	0.48	0.45	1.14	1.03–1.26	6.8×10^{-3}
Meta-analysis					1.11	1.07–1.15	2.4×10^{-8}
Replication							
Australian #1	145	116	0.32	0.22	1.66	1.16–2.38	5.8×10^{-3}
Australian #2	431	567	0.27	0.24	1.22	1.00–1.48	6.2×10^{-2}
Combined	576	683	0.29	0.23	1.32	1.11–1.58	1.7×10^{-3}

Cont, control; OR, odds ratio. The allele frequency is for the C allele. Note that the European results show the raw allele frequencies across cohorts, with the OR calculated from logistic regression that includes covariates

more common in Chinese than in Europeans (0.46 vs. 0.25). The association result was replicated in an independent Australian sample (145 cases, 116 controls, OR = 1.66; 95% CI: 1.16–2.38; $p = 5.8 \times 10^{-3}$) and had the same direction of effect in a second Australian sample (431 cases, 567 controls, OR = 1.22; 95% CI: 1.00–1.48; $p = 6.2 \times 10^{-2}$), giving a combined replication OR of 1.32 (95% CI: 1.11–1.58; $p = 1.7 \times 10^{-3}$) (Table 1).

Functional relevance of GPX3 and TNIP1. Both *GPX3* and *TNIP1* are genes that could have functional relevance for ALS. The protein glutathione peroxidase 3 (GPX3), is an antioxidant molecule functionally related to superoxide dismutase 1 (SOD1)[13]; many *SOD1* single-nucleotide variants are pathogenic for ALS. In a mass spectrometric screen of sera of $SOD1^{H46R}$ rats compared to their wild-type (WT) controls in the pre-symptomatic stage (12 weeks of age) of ALS, Gpx3 was detected as one of the two significant results (1.3-fold increase in expression)[14]. In the same study, Gpx3 expression was significantly lower (0.74 fold, $p = 0.009$) compared to WT controls by disease end stage, a finding which was replicated in blood sera of sporadic ALS cases ($n = 18$) and controls ($n = 35$) (GPX3 0.41-fold lower, $p = 0.008$)[14]. Both *GPX3* and *TNIP1* are functionally associated with NF-κB, the master regulator of inflammation[17, 19], with upregulation of NF-κB associated with death of motor neurones[15]. Protein–protein interaction analysis[18] links GPX3 to SOD1 and TNIP1 to OPTN, and *OPTN* also harbours mutations associated with familial ALS[5]. *TNIP1* is associated with a wide range of immune disorders[22, 23], although our most associated SNP (rs10463311) is not in LD with specific SNPs associated with these disorders[24]. We investigated differential expression of *GPX3* and *TNIP1* between ALS patients and controls, but given small sample sizes, the results were not conclusive (Supplementary Note 1, Supplementary Table 3, Supplementary Fig. 6). In a pleiotropy informed analysis[25] applied to the European GWAS summary statistics[7], rs10463311 was identified as an ALS-associated SNP, providing additional, albeit not fully independent, support for this locus.

Gene-based association analysis. No genes were significantly associated with ALS from gene-based association analysis implemented in fastBAT[26] of Chinese data (based on Bonferroni correction for ~18,000 autosomal genes, significance declared at 2.8×10^{-6}), but meta-analysed results (Supplementary Table 4) identified multiple genes (reflecting LD and overlapping gene boundaries) at the previously reported chromosome 5, 9, 14 and 17 GWAS loci. Two new loci on Chromosome 17 (17q12 and 17q21.2) were also significant (minimum genic $p = 3.3 \times 10^{-7}$ and 1.2×10^{-7}, respectively). The former locus was also supported by summary statistic-based Mendelian randomization (SMR) analysis[27] that combines the disease–SNP association with gene expression–SNP association results and has GW significance threshold of $p_{SMR} < 8.4 \times 10^{-6}$) (Supplementary Fig. 7; Supplementary Data 2), with most significant association for *GGNBP2* (European only $p_{SMR} = 4.6 \times 10^{-6}$; meta-analysis $p_{SMR} = 9.8 \times 10^{-6}$). The two replication samples did not provide support for the *GGNBP2* SNP implicated from the SMR analysis (Supplementary Table 5); larger sample sizes are needed to confirm the association and to provide evidence to exclude *ZNHIT3* ($p_{SMR} = 3.1 \times 10^{-5}$) or *MYO19* ($p_{SMR} = 2.2 \times 10^{-4}$) as contributing to the association in this region. Gene-set pathway analysis implemented in MAGMA and applied to the Chinese/European meta-analysis results did not find any ALS significant pathways that passed multiple testing correction (Supplementary Table 6).

Discussion

In summary, using a cross-ethnic design we identify association of the *GPX3-TNIP1* locus with ALS. This locus was identified by combining GWAS results from our Chinese data with the largest European GWAS data[7] and replicated in independent Australian samples. In addition, *GGNBP2* was identified using gene-based analysis and SMR analysis, although further replication is needed to confirm this result. The discovery of a novel risk locus significantly advances our understanding of ALS aetiology.

Methods

Chinese ALS cases and controls. The samples comprised 1,324 ALS cases and 3,115 controls. ALS cases were recruited from the Department of Neurology, the Peking University Third Hospital (Beijing, China) from 2003 to 2013. The cases were diagnosed by a neurologist specializing in ALS using the revised El Escorial criteria[28]. The controls are individuals who attended the Peking University Third Hospital, Peking University Sixth Hospital or Shanghai Changzheng Hospital (Shanghai) with no medical or family history of neurological disorders. All cases and controls are of Chinese origin from Mainland China and provided written informed consent for the study. The sample collections were approved by the ethics committees at the respective hospitals[12]. The study is compliant with the Guidance of the Ministry of Science and Technology (MOST) for the Review and Approval of Human Genetic Resources. Analyses conducted at the University of Queensland were approved by the University human research ethics committee.

Australian replication cohort 1. ALS cases were recruited from the Royal Brisbane & Women's Hospital (RBWH), Brisbane, Queensland and the Macquarie University Multidisciplinary Motor Neurone Disease Clinic[29], New South Wales. The cases ($N = 159$) were diagnosed using the revised El Escorial criteria[10]. The controls are healthy individuals ($N = 132$), sourced from 4 different sites, RBWH (27 individuals), Neurology at Macquarie University, Sydney (25 individuals), the Older Australian Twin Study (OATS)[30] comprising 90 monozygotic (MZ) twin pairs recruited in Brisbane (QIMR Berghofer Medical Research Institute (QIMR)) and Sydney (University of New South Wales (UNSW)) and Melbourne (University of Melbourne (UM)). The OATS study recruits MZ twins aged ≥65 years and were chosen for this study because the Discovery sample controls were younger than Discovery sample cases. Twin pair data helped in quality control checks but only one twin from each pair was used in analyses. The subjects provided written informed consent for the study. The study was approved by the RBWH[31], QIMR, UNSW, UM, University of Queensland and Macquarie University Research Ethics Committees.

Australian replication cohort 2. Patients and controls were ascertained from Macquarie University Multidisciplinary Motor Neurone Disease and Neurology Clinics, Sydney and from the Australian MND DNA bank. Patients were diagnosed with definite or probable ALS according to the revised El Escorial criteria. Patients with a family history for ALS were excluded. Control subjects were healthy individuals free of neuromuscular diseases. DNA from 471 cases and 586 controls were available for genotyping. The subjects provided written informed consent for the study. The study was approved by Macquarie University Research Ethics Committee.

DNA extraction. In the Chinese cohort, genomic DNA was extracted from whole blood using the DNA Extraction Kit (Beijing Aide Lai Biotechnology Co. Ltd., Beijing, China). In the Australian replication cohorts, the majority of DNA was extracted from fresh whole blood using manual extraction protocols, except for 90% (118 out of 131) of UNSW/UM control samples, where DNA was extracted from frozen whole blood or lymphocytes using an automated purification system, Qiagen Autopure LS (Qiagen, Valencia, CA, USA).

Genome-wide association study. We performed GW genotyping in the discovery cohort using the Illumina HumanOmni ZhongHua-8 v1.0 and v1.1 arrays. These arrays contain 900,015 (v1.0) and 894,517 (v1.1) variants, respectively. Before testing for the association between each variant and disease status, we carried out quality control (QC) steps to identify and exclude poor quality samples and genetic variants. We excluded individuals based on the following QC filters: (i) genotyping call rate <99% (134 individuals); (ii) sex mismatch between genotype and clinical information (6 individuals); (iii) ancestry outliers (6 SDs from HapMap-CHB means of PC1 and PC2; 30 individuals); and (iv) duplicated or related individuals (genetic relationship matrix >0.05; 195 individuals). We excluded genetic variants based on the following criteria: (i) low genotype call rate <99%; (ii) MAF <1%; (iii) deviation from Hardy–Weinberg equilibrium $p < 10^{-6}$; and (iv) differential missingness in genotypes between cases and controls ($p < 10^{-6}$). After these QC steps, 1,234 cases and 2,850 controls with genotypic information from 753,038 markers remained for the subsequent analyses.

We imputed unobserved genotypes into the 1000 Genomes Project Phase 1 v3 (all ethnicities) using samples and markers that passed QC. We implemented a two-step process, i.e., haplotyping using HAPI-UR[32] and imputation using IMPUTE[33]. We imputed 38,033,906 SNPs, but after QC (i.e., excluding markers with MAF <0.01, imputation quality score <0.80 and HWE $p < 10^{-6}$), 6,613,544 SNPs were available for analysis.

Validation sample genotyping. The first validation sample was genotyped on the Illumina Human Core Exome Array. QC and imputation followed the same pipeline as for the Chinese samples. After QC, 145 cases and 116 controls were available for analysis. For the second validation sample, SNPs were genotyped via Taqman assay such that the reaction mix included 1.0 µl of genomic DNA (10 ng/µl), 0.25 µl Custom TaqMan genotyping assay 20× (Life Technologies), 2.5 µl TaqMan SNP genotyping MasterMix 2X (Life Technologies) and 6.25 µl MilliQ. The thermocycler program included 30 s at 60 °C, 10 min at 95 °C, followed by 40 cycles of 15 s at 95 °C and 1 min at 60 °C and a final step of 30 s at 60 °C. Fluorescent signals were analysed on a Viia7 Real-Time PCR System and genotypes were determined by allelic discrimination using the Viia7 Real-Time PCR System Software (Life Technologies). Genotype calling rates were 94% for rs4958872 (LD $r^2 = 1$ proxy for rs10463311) and 91% for rs9906189. After QC, 431 cases and 567 controls were available for analysis.

Genetic association analysis. The association analysis between genetic variants and disease was conducted using a linear mixed model framework implemented in GCTA (mlma-loco)[34]. To compare the results, we also used a logistic regression model by fitting five principal components as covariates. Genomic inflation factor was calculated as the median of Chi-square test statistics divided by its expected value (0.455).

Gene-based analysis. To test for the association between a set of variants within a gene (±50 kb) and ALS, we used GCTA-fastBAT[26] with SNP association analysis *p* values as input. This test complements SNP–disease association analysis, identifying genes that may show evidence for independent associations that individually have not achieved association significance. For Chinese data analysis, we used our own GWAS data as the reference to calculate LD and ARIC samples (dbGAP accession phs000090.v1.p1) for the European sample.

Whole-genome estimation analysis. Genomic relationship matrix (GRM) restricted maximum likelihood (GREML) analysis using GCTA[19, 35, 36] was used to estimate the total contribution of common genetic variants on the liability of ALS or SNP-heritability. This analysis fits all SNPs simultaneously in a mixed model linear framework to estimate the proportion of variance in disease liability explained by all SNPs. To avoid bias, for example, due to common environmental factors, we excluded related individuals based on GRM values >0.05. Lifetime disease risk of 0.002 was used in the conversion of the estimate to the liability scale[37] (compared to 0.0025 used in the European conversion, although the results are robust to these choices). LD-score regression[20] was applied to GWAS summary

statistics as an alternative method to estimate the contribution of common genetic variants to variation in the liability of ALS.

Genetic overlap analysis. We considered estimation of the genetic correlation between ALS risk in Europeans and Chinese, using popcorn[38] (the cross-ethnicity LDscore regression method), but calculated[39] that the relatively small sample size for the Chinese cohort would generate an unacceptably large SE. Instead we used polygenic risk scoring (PRS) to investigate the genetic relationship between ALS in the two ethnicities. PRS were estimated for all Chinese cases and controls as the sum of risk alleles weighted by the log OR of association estimated in the European GWAS. Eight PRS were constructed for each individual using independent SNPs (based on SNPs pruned ($r^2 < 0.25$ in 200-kb window) that are significant at p value thresholds of 0.001, 0.005, 0.01, 0.05, 0.10, 0.25, 0.5 and 1. We also constructed a PRS using all SNPs without pruning for LD because of the difference in allele frequencies and LD between ethnicities. Association between the case–control status and PRS was evaluated by logistic regression. Binomial sign tests were also used to evaluate evidence of overlap in signal between Chinese and European association statistics.

Meta-analysis. Inverse variance meta-analysis was conducted between the largest GWAS for ALS in European[7] and our Chinese GWAS results using METAL[40].

In silico functional analyses. To help interpret biological function of the SNP– and gene–ALS associations, gene-set pathway analyses were performed using MAGMA[41]; this method was selected based on results of a method comparison study[42]. Gene-set pathway analyses aim to identify sets of biological pathways that are relevant to disease based on a set of disease-associated variants[42].We also conducted SMR analysis[27] that combines the GWAS summary statistics with gene expression association results. Here we used gene expression from blood[43] as this is currently the largest gene expression quantitative trait loci data set. The SMR test identifies pleiotropic association of a variant that affects both the expression level of a gene and the trait. The SMR-HEIDI test attempts to determine whether the effect of the disease-associated gene on gene expression reflects a single causal variant, thus prioritizing loci for functional follow-up studies.

Data availability. GWAS summary statistics results and gene expression data are available from http://cnsgenomics.com/data/benyamin_et_al_2017_nc/BenyaminEtAl_NatComm_Data.zip.

Received: 11 January 2017 Accepted: 30 June 2017
Published online: 20 September 2017

References

1. Chiò, A. et al. Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology* **41**, 118–30 (2013).
2. Johnston, C. A. et al. Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J. Neurol.* **253**, 1642–3 (2006).
3. Kiernan, M. C. et al. Amyotrophic lateral sclerosis. *Lancet* **377**, 942–955 (2011).
4. Chio, A. et al. Prognostic factors in ALS: A critical review. *Amyotroph. Lateral Scler.* **10**, 310–323 (2009).
5. Renton, A. E. et al. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17–23 (2014).
6. Cirulli, E. T. et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436–1441 (2015).
7. van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
8. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
9. Kenna, K. P. et al. NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1037–1042 (2016).
10. Liu, M. S., Cui, L. Y. & Fan, D. S., Chinese ALS Association. Age at onset of amyotrophic lateral sclerosis in China. *Acta Neurol. Scand.* **129**, 163–7 (2014).
11. Liu, Q. et al. Mutation spectrum of Chinese patients with familial and sporadic amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry* **87**, 1272–1274 (2016).
12. He, J. et al. C9orf72 hexanucleotide repeat expansions in Chinese sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging* **36**, 2660.e1–2660.e8 (2015).
13. Chi, L., Ke, Y., Luo, C., Gozal, D. & Liu, R. Depletion of reduced glutathione enhances motor neuron degeneration in vitro and in vivo. *Neuroscience* **144**, 991–1003 (2007).
14. Tanaka, H. et al. ITIH4 and Gpx3 are potential biomarkers for amyotrophic lateral sclerosis. *J. Neurol.* **260**, 1782–97 (2013).
15. Frakes, A. E. et al. Microglia induce motor neuron death via the classical NF-κB pathway in amyotrophic lateral sclerosis. *Neuron* **81**, 1009–23 (2014).
16. Oliveira-Marques, V., Marinho, H. S., Cyrne, L. & Antunes, F. Role of hydrogen peroxide in NF-kappaB activation: from inducer to modulator. *Antioxid. Redox Signal.* **11**, 2223–43 (2009).
17. Rahighi, S. et al. Specific recognition of linear ubiquitin chains by NEMO is important for NF-kappaB activation. *Cell* **136**, 1098–109 (2009).
18. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
19. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–9 (2010).
20. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–5 (2015).
21. Deng, M. et al. Genome-wide association analyses in Han Chinese identify two new susceptibility loci for amyotrophic lateral sclerosis. *Nat. Genet.* **45**, 697–700 (2013).
22. Gateva, V. et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1228–33 (2009).
23. Nair, R. P. et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.* **41**, 199–204 (2009).
24. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
25. McLaughlin, R. L. et al. Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat. Commun.* **8**, 14774 (2017).
26. Bakshi, A. et al. Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci. Rep.* **6**, 32894 (2016).
27. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–7 (2016).
28. Brooks, B. R., Miller, R. G., Swash, M. & Munsat, T. L. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Other Motor Neuron Disord* **1**, 293–9 (2000).
29. Williams, K. L. et al. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nat. Commun.* **7**, 11253 (2016).
30. Sachdev, P. S. et al. A comprehensive neuropsychiatric study of elderly twins: the Older Australian Twins Study. *Twin Res. Hum. Genet.* **12**, 573–82 (2009).
31. Devine, M. S., Kiernan, M. C., Heggie, S., McCombe, P. A. & Henderson, R. D. Study of motor asymmetry in ALS indicates an effect of limb dominance on onset and spread of weakness, and an important role for upper motor neurons. *Amyotroph. Lateral Scler. Frontotemporal Degener* **15**, 481–7 (2014).
32. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91**, 238–51 (2012).
33. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–13 (2007).
34. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–6 (2014).
35. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
36. Yang, J. et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–25 (2011).
37. Lee, S. H. et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
38. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
39. Visscher, P. M. et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
40. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
41. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
42. de Leeuw, C. A., Neale, B. M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **17**, 353–364 (2016).
43. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–43 (2013).
44. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).

Acknowledgements

We thank all the participants for their generous contribution to this study. We also thank staff and researchers at the Peking University Third Hospital (Beijing, China), the Royal Brisbane and Women Hospital (Brisbane, Australia) and Multidisciplinary Motor Neurone Disease Clinic (Sydney, Australia). This work was funded by the Australian Research Council (ARC) Linkage Grant (M.A.B., P.F.B., P.M.V., H.X., R.J.W., B.J.M., D.C.R. and M.M), the Peter Goodenough Foundation, the National Natural Science Foundation of China (grants to D.S.F.: 81030019, and J.H.: 81601105 and Z.B.J.: 81522014), the National Health and Medical Research Council (NHMRC) (N.R.W.: 1078901, 1083187, B.B.: 1084417, 1079583, P.M.V.: 1078037, I.P.B.: 1095215, K.L.W.:

1092023), MNDRIA (P.G.N.: Grant-in-Aid 2013, B.B.: Mick Roger Benalla Grant, F.G.: Bill Gole Fellowship, I.P.B.: MND Australia Leadership Grant, E.P.M. and J.A.F.: scholarship top ups), Ross Maclean Senior Research Fellowships (M.M, R.J.W.), the Sylvia & Charles Viertel Charitable Foundation (J.Y.). O.A.T.S. was supported by an NHMRC/ARC Strategic Award (Grant 401162) and the Project Grant (104532) and was facilitated through access to the Australian Twin Registry, which is funded by an NHMRC Enabling Grant (310667).

Author contributions

M.A.B., D.F., R.H.W., H.X. and P.F.B. conceived the study. M.A.B., P.F.B., P.M.V., H.X., R.J.W., D.C.R., B.J.M., M.M., R.H.W., N.R.W., B.B., F.G. and P.G.N. obtained funding. B.B., J. He, Q.Z., J.G., P.L., Z.L., Z.-H.Z., P.M.V. and N.R.W. performed statistical analysis. B.B., F.G., Z.L., M.M., S. Shah, J.Y. and N.R.W. performed follow-up analysis. D.F., J. He, L.T., L.C., X.L. and M.A.B. provided Chinese ALS and control samples. H.X., D.Z., W.Y., X.W., T.L., M.L. and M.A.B. provided Chinese GWAS control samples. J.W., Z.-B. J., Z. Li, A.K.H., J.Y., P.G.N., R.L.J., M.D., S.F., S.T.N. and T.J.B. provided data. H.-W.D., Y.L., S.R., Y.-Y.H., L.J.T. and X.-D.C. provided HNU WES controls. P.A.M., R.D.H., R.P., D.B.R., I.P.B., P.S., K.L.W., A.K.H., E.P.M., J.A.F. and N.R.W. provided Australian ALS and control samples. L.A., K.C., J.E., J. Harris, S. Song, A.K.H., L.W., H.X. and M.A.B. performed genotyping. J.H.V., W.v.R., A.A.-C. and L.H.v.d.B. provided European GWAS data. B.B., Q.Z., J.G., F.G. and N.R.W. wrote the first draft of the manuscript. All authors contributed to revision of the manuscript.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00471-1.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Beben Benyamin[1,2], Ji He[3], Qiongyi Zhao[1], Jacob Gratten[1,2], Fleur Garton[1,2], Paul J. Leo[4,5], Zhijun Liu[1,2], Marie Mangelsdorf[1], Ammar Al-Chalabi[6], Lisa Anderson[4,5], Timothy J. Butler[1], Lu Chen[3], Xiang-Ding Chen[7], Katie Cremin[4,5], Hong-Weng Deng[8], Matthew Devine[9], Janette Edson[1], Jennifer A. Fifita[10], Sarah Furlong[1], Ying-Ying Han[11], Jessica Harris[4,5], Anjali K. Henders[1,2], Rosalind L. Jeffree[12], Zi-Bing Jin[13], Zhongshan Li[14], Ting Li[15], Mengmeng Li[15], Yong Lin[11], Xiaolu Liu[3], Mhairi Marshall[4,5], Emily P. McCann[10], Bryan J. Mowry[1], Shyuan T. Ngo[1,16], Roger Pamphlett[17], Shu Ran[11], David C. Reutens[18], Dominic B. Rowe[19], Perminder Sachdev[20,21], Sonia Shah[1], Sharon Song[4,5], Li-Jun Tan[7], Lu Tang[3], Leonard H. van den Berg[22], Wouter van Rheenen[22], Jan H. Veldink[22], Robyn H. Wallace[1], Lawrie Wheeler[4,5], Kelly L. Williams[10], Jinyu Wu[14], Xin Wu[15], Jian Yang[1,2], Weihua Yue[23,24], Zong-Hong Zhang[1], Dai Zhang[23,24], Peter G. Noakes[9], Ian P. Blair[10], Robert D. Henderson[1,9], Pamela A. McCombe[9,25], Peter M. Visscher[1,2], Huji Xu[15], Perry F. Bartlett[1], Matthew A. Brown[4,5], Naomi R. Wray[1,2] & Dongsheng Fan[3]

[1]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia. [2]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia. [3]Department of Neurology, Peking University, Third Hospital, No. 49, North Garden Road, Haidian District, 100191 Beijing, China. [4]University of Queensland Diamantina Institute, The University of Queensland, Translational Research Institute, Brisbane, Queensland 4102, Australia. [5]Institute of Health and Biomedical Innovation, Queensland University of Technology, Translational Research Institute, Brisbane, Queensland 4102, Australia. [6]Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London, WC2 R2LS, UK. [7]Laboratory of Molecular and Statistical Genetics and State Key Laboratory of Developmental Biology of Freshwater Fish, College of Life Sciences, Hunan Normal University, Changsha, 410081 Hunan, China. [8]Department of Global Biostatistics and Data Science, School of Public Health and Tropical Medicine, Center for Bioinformatics and Genomics, Tulane University, 1440 Canal St, Suite 2001, New Orleans, LA 70112, USA. [9]Department of Neurology, Royal Brisbane & Women's Hospital, Brisbane, Queensland 4029, Australia. [10]Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [11]Center of System Biomedical Sciences, University of Shanghai for Science and Technology, 334, Jungong Road, Yangpu District, 200093 Shanghai, China. [12]Kenneth G. Jamieson Department of Neurosurgery, Royal Brisbane & Women's Hospital, Herston, Queensland 4029, Australia. [13]Division of Ophthalmic Genetics, Laboratory for Stem Cell and Retinal Regeneration, The Eye Hospital of Wenzhou Medical University, 325027 Wenzhou, China. [14]Institute of Genomic Medicine, Wenzhou Medical University, 325027 Wenzhou, China. [15]Department of Rheumatology and Immunology, Shanghai Changzheng Hospital, The Second Military Medical University, 200003 Shanghai, China. [16]School of Biomedical Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia. [17]Stacey MND Laboratory, Discipline of Pathology, Brain and Mind Centre, The University of Sydney, Sydney, New South Wales 2006, Australia. [18]The Centre for Advanced Imaging, The University of Queensland, Brisbane, Queensland 4072, Australia. [19]Department of Medicine, Faculty of Medicine and Health Sciences, Multidisciplinary Motor Neurone Disease Clinic, Macquarie University, Sydney, New South Wales 2109, Australia. [20]Centre for Healthy Brain Ageing, School of Psychiatry, Faculty of Medicine, The University of New South Wales, Sydney, New South Wales 2052, Australia. [21]Neuropsychiatric Institute, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [22]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands. [23]Institute of Mental Health, The Sixth Hospital, Peking University, 100191 Beijing, China. [24]Key Laboratory of

Mental Health, Ministry of Health & National Clinical Research Center for Mental Disorders, Peking University, 100191 Beijing, China. [25]UQ Centre for Clinical Research, The University of Queensland, Royal Brisbane & Women's Hospital, Brisbane, Queensland 4029, Australia. Beben Benyamin and Ji He contributed equally to this work. Matthew A. Brown, Naomi R. Wray and Dongsheng Fan jointly supervised this work.

Fifita, J. A., et al. (2017). Genetic and Pathological Assessment of hnRNPA1, hnRNPA2/B1, and hnRNPA3 in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Neurodegenerative Diseases*, 17, p. 304-312.

DOI: 10.1159/000481258

Monozygotic twins and triplets discordant for amyotrophic lateral sclerosis display differential methylation and gene expression

Ingrid S. Tarr[1], Emily P. McCann[1], Beben Benyamin[2,3], Timothy J. Peters[4], Natalie A. Twine[5], Katharine Y. Zhang[1], Qiongyi Zhao[2], Zong-Hong Zhang[2], Dominic B. Rowe[6], Garth A. Nicholson[7,8], Denis Bauer[5], Susan J. Clark[4,9], Ian P. Blair[1] and Kelly L. Williams[1]*

[1] Centre for Motor Neuron Disease Research, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia

[2] Queensland Brain Institute, University of Queensland, Queensland, Australia

[3] Australian Centre for Precision Health, University of South Australia Cancer Research Institute, School of Health Sciences, University of South Australia, Adelaide, Australia

[4] Epigenetics Research Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia

[5] Health and Biosecurity Business Unit, Commonwealth Scientific and Industrial Research Organisation, Sydney, New South Wales, Australia

[6] Department of Clinical Medicine, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, New South Wales, Australia

[7] ANZAC Research Institute, University of Sydney, Sydney, New South Wales, Australia

[8] Molecular Medicine Laboratory, Concord Hospital, Sydney, New South Wales, Australia

[9] St Vincent's Clinical School, UNSW Sydney, 2010, New South Wales, Australia.

***Corresponding author:**

Dr Kelly Williams

Macquarie University Centre for Motor Neuron Disease Research

Faculty of Medicine and Health Sciences
Level 1, 75 Talavera Road
Macquarie University, NSW 2109
Ph: +612 9850 2731
Email: kelly.williams@mq.edu.au

Keywords
Amyotrophic lateral sclerosis
SOD1
C9orf72
Twin
Triplet
Discordant
DNA methylation
Transcriptome

Abstract

Background

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by the loss of upper and lower motor neurons. ALS exhibits high phenotypic variability including age and site of onset, and disease duration. Gene mutations, which account for a small proportion of cases, can also show variable penetrance. Together, this strongly implicates modifying factors including those that impact gene expression.

Disease discordant monozygotic (MZ) twins/triplets provide a unique opportunity to uncover epigenetic and transcriptomic factors that may modify a phenotype and avoid confounding factors such as genetic variation and early developmental environment. A cohort of Australian monozygotic twins (n=3 pairs) and triplets (n=1 set) were recruited that are discordant for ALS and represent sporadic ALS and the two most common types of familial ALS, linked to *C9orf72* and *SOD1*. We sought to identify longitudinally consistent modifying factors by examining whole blood-derived differential DNA methylation and gene expression.

Results

Longitudinal differentially methylated genes were mostly unique to a single twin/triplet set, yet a small group of genes were differentially methylated across twin/triplet sets and showed differential expression profiles. Two of these genes, *RAD9B* and *C8orf46*, showed significant differential methylation in a validation cohort of >1000 ALS cases and controls ($p = 2.5E-5$ and $p = 0.049$ respectively). Combined longitudinal methylation-transcription analysis within

a single twin set implicated *CCNF*, *DPP6*, *RAMP3*, and *CCS*, which have been previously associated with ALS.

Gene Ontology analysis of longitudinal transcriptome data implicated up to 8-fold enrichment (FE) of genes associated with immune function pathways (p = 1.4E-4) and under-representation of transcription and protein modification genes (FE = 0.2, p = 0.01) in sporadic ALS. DNA methylation indicated that increased methylation age is a signature of ALS in older patients (p = 1.3E-5).

Conclusions

Analysis of cytosine methylation and mRNA transcription in ALS-discordant monozygotic twins/triplets identified consistent longitudinal differential DNA methylation and gene expression. Validation of these changes in a large Australian sporadic ALS suggest a broader role in ALS. Differentially methylated and expressed genes and their functional pathways may contribute to variable disease penetrance and offer targets for therapeutic development.

Background

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by the rapidly progressive loss of the upper and lower motor neurons. Disease onset commonly occurs in middle to late age [1] and typically results in death within three to five years. Existing treatments are of limited effect, and despite intensive effort, the pathogenic mechanisms underlying disease are still poorly understood. A recognised family history

(familial ALS) is seen in approximately 10% of cases while the remainder are considered sporadic [2]. The familial and sporadic forms of the disease are clinically and pathologically indistinguishable [3].To date, the only proven cause of ALS are gene mutations leading to motor neuron death. Pathogenic repeat expansions in the *C9orf72* gene and missense mutations in the *SOD1* gene are the most frequent known causes of ALS worldwide, yet no cause has been identified for the majority of patients (>80%, [4]). Even in those individuals with a proven causal gene mutation, inter- and intra-familial phenotypic heterogeneity is commonly observed [5, 6]. Age of disease onset may vary by more than 60 years and disease duration may be measured in months or in decades. Affected individuals, particularly those with a *C9orf72* repeat expansion, may present with ALS or frontotemporal dementia (FTD), or a mixed phenotype. Causal mutations may show incomplete penetrance [5] and indeed monozygotic twins are more commonly discordant for ALS than concordant [7]. Taken together, this phenotypic variability suggests a significant contribution from modifying factors in disease manifestation.

Epigenetic and transcriptional profiling have implicated differential DNA methylation and/or gene expression in ALS. *C9orf72* has been shown to have increased methylation [8, 9] and decreased transcription [10, 11] in ALS/FTD patients with the pathogenic repeat expansion. Other major ALS genes, however, including *SOD1*, *FUS* and *TARDBP*, are generally unmethylated and show no differences between patients and controls [12-14]. Nevertheless, changes in expression of some ALS genes is apparent in sporadic disease [15]. Whole methylome and transcriptome studies in spinal cord and blood tissue have found global changes [12, 16, 17] and implicated various genes, pathways and several overlapping themes including changes that affect immune response [16, 18, 19] and cellular transport [20, 21].

Disease discordant monozygotic (MZ) twins hold great potential for studies that seek to identify epigenetic and transcriptomic factors that modify the phenotype of complex human diseases. Identical twin studies can account for confounding factors such as genetic variation and the early development environment. Such studies have informed understanding of phenotypic variation in Parkinson's disease [22], Alzheimer's disease[23], systemic lupus erythematosus[24], and depression [25], among others. Previous DNA methylation studies of known causal ALS genes in ALS-discordant MZ twins found no aberrant methylation between twins [26, 27], while twin-based methylome-wide studies suggested a different epigenetic age in affected twins [27, 28] and identified potentially altered GABA signalling [27] and immune response [29]. Nevertheless, further studies are required because the differentially methylated sites implicated in initial screens have often failed to be validated in targeted studies using bisulphite pyrosequencing [28]. Similarly, candidate molecular pathways have shown limited overlap between twin sets [27] and changes in methylation are yet to be linked to changes in transcription. It remains unclear which of the observed differences in either DNA methylation or gene expression reflect ALS discordance between co-twins. It is also unclear whether these differences in DNA methylation correlate with differential gene expression on a transcriptome-wide scale.

In this study, we undertook comprehensive methylome- and transcriptome-wide analysis of a longitudinal ALS-discordant cohort comprising MZ triplets and twins, representing the three most common types of ALS, *C9orf72*-linked ALS, *SOD1*-linked ALS and sporadic ALS. We analysed methylome- and transcriptome-wide data, independently and in combination, in an attempt to identify disease-relevant methylation changes and their downstream impact. Co-twin analyses indicated a significant interaction effect between age and disease status on DNA methylation age, with older twins showing a consistent difference between ALS-affected and unaffected co-twins in a longitudinal series. Furthermore, we

identified several genes likely to contribute to ALS through integration of longitudinal twin genome-wide DNA methylation and transcription data, further assessed in a large sporadic ALS case-control cohort.

Results

ALS-discordant and control twin/triplet sets

Clinical and sample information for the three discordant MZ twin sets, one discordant MZ triplet set and two control twin sets are included in table 1. Pedigrees and extended pedigrees are shown in figure 1. All individuals with ALS have been screened for causal mutations in known ALS genes. The FALS twin set has a pathogenic hexanucleotide repeat expansion in *C9orf72*. The FALS triplet set harbours a *SOD1* p.I114T mutation.

Table 1. Twin cohort details

MZ set	**ALS**	**Status**	**Sex**	**Mutation**	**Age of onset**	**Age at sampling[A]**	**Duration (months)**	**450K samples (n)[B]**	**RNA-Seq samples (n)**	**EpiTYPER samples (n)**
Female SALS twin set	SALS	ALS	F		42.7	43.5 - 45.1	Alive at 51 months	8 (+1)	-	-
		Unaffected	F			43.9 - 45.1		8 (+1)	-	-
Male SALS twin set	SALS	ALS	M		78.5	79.8 - 80.2	28.4	3 (+1)	3	-
		Unaffected	M			79.8 - 80.2		3 (+1)	3	-

***C9orf72* twin set**	FALS	ALS	M	*C9orf72 HRE*	52	54.1	36	1	-	1
		Asymptomatic	M	*C9orf72 HRE*		54.3 – 55		2	-	2
***SOD1* triplet set**	FALS	ALS	F	*SOD1* p.I114T	50	50.3	Unknown	1	-	1
		Asymptomatic	F	*SOD1* p.I114T		50.3		1	-	1
		Asymptomatic	F	*SOD1* p.I114T		50.3		1	-	1
Control twin set 1	NA	Control	F		NA	46.1	NA	1	-	-
		Control	F			46.1		1	-	
Control twin set 2	NA	Control	M		NA	36.8	NA	-	-	1
		Control	M			31.8 - 43.0[C]				3[C]

HRE: hexanucleotide repeat expansion

[A] Presence of an age range indicates longitudinal samples were collected

[B] Number of technical replicates during blood collection indicated in brackets

[C] Middle sample matched to co-twin

Targeted analysis of methylation in mutation-known MZ sets

To assess whether differential methylation of the *C9orf72* or *SOD1* CpG islands were associated with the disease discordance we observe in the *C9orf72* twin set and *SOD1* triplets, we investigated the status of CpG methylation of the *C9orf72* and *SOD1* CpG islands. To perform a high-density, targeted analysis, we used EpiTYPER, with additional

support from a number of Infinium HumanMethylation450K CpG sites present in the same region. Due to the small sample sizes available, results are descriptive only.

SOD1 methylation in the _SOD1_ MZ triplet set shows a consistent methylation pattern

We used EpiTYPER to quantify methylation of the upstream *SOD1* CpG island encompassing the *SOD1* promoter region and exon 1 in the discordant MZ triplets carrying the *SOD1* p.I114T mutation and a pair of control twins from another SOD1 p.I114T family that were negative for *SOD1* mutation. Additionally, five *SOD1* CpG sites present in the Infinium HumanMethylation450K data set were located within the CpG island (fig. 2A). Neither the 23 CpG units within the CpG island, nor the five 450K *SOD1* CpG sites, showed any consistent methylation differences between ALS affected and ALS unaffected MZ triplets, nor control twins. (fig. 2A).

No differences were observed in _C9orf72_ methylation in the _C9orf72_ MZ twin set

The quantitative methylation status of two CpG islands associated with *C9orf72* was determined using EpiTYPER. The amplicons covered the entirety of both CpG islands, the promoter region and adjacent intronic/intergenic regions. The intronic pathogenic $(GGGGCC)_n$ repeat expansion (indicated with a black diamond in fig. 2B) is flanked by the two CpG islands. In the disease discordant FALS twin set harbouring a *C9orf72* expansion, methylation across 28 CpG sites measured by the EpiTYPER assay are highly concordant and generally unmethylated (fig. 2B). Similarly, in the four 450K probes associated with *C9orf72*, none of the CpG sites show a clear difference in methylation between the co-twins (fig. 2B).

Whole methylome analysis of disease discordant MZ twins/triplets

Co-twin/triplet differences in DNA methylation age reflects an age-dependant effect

Horvath's [30] DNA methylation age algorithm was used to determine epigenetic age from the methylation signature of each twin/triplet sample in table 1. We tested the association of methylation age with disease status and chronological age in a mixed model while controlling for sex. The effect of disease status on methylation age was found to be highly dependent upon chronological age (p=1.3E-5, fig. 3A). Briefly, with increasing age, asymptomatic co-twins were estimated to have a younger epigenetic age than their ALS-affected twin. This result was most evident in the approximately 20-year difference in methylation age between twins in the oldest disease discordant twin set of this cohort (fig. 3A).

Global methylation and cell type proportions do not show any effect of disease status

Global methylation was calculated as the mean methylation across all Infinium HumanMethylation450K CpG sites passing data processing (n=386,183). No significant effect of disease on global methylation was found when controlling for sex and age at sample collection (p=0.08, fig. 3B). To better reflect influence on transcription, CpG sites were classified according to CpG density: high density CpG islands, intermediate density in islands, island shores, and low CpG density. Mean methylation within each of these four levels of CpG density does not show any effect of disease status (HC, p=0.93; IC, p=0.99; ICshore, p=0.82; LC, p=0.093, fig. S2).

Blood cell proportions for each twin/triplet sample were estimated using Houseman et al.'s algorithm [31] and the six cell types were assessed for association with disease status. Disease status did not have a significant effect on any of the cell types when controlling for age at sample collection and sex (all $p > 0.2$, fig. 3C).

Differentially methylated probes were identified across discordant MZ twins/triplets

Statistical significance and the magnitude of pairwise methylation differences were combined to detect differentially methylated probes in discordant MZ twin studies. 59 probes were identified as differentially methylated across twin/triplet sets (full list in table S5, 9 top-ranked probes shown in fig. 4A). All 59 probes were used for hierarchical clustering and PCA of the longitudinal MZ cohort to investigate the presence of a disease signature. Both hierarchical clustering and PCA did not indicate that samples cluster by disease status, but rather approximately by twin set and individual, where longitudinal samples were available (fig. 4B, C).

The 59 probes were subsequently investigated in our large case-control 450K methylation validation data set. After FDR correction, 2 of the 59 probes showed significantly differential methylation between cases and controls when controlling for age and sex (*RAD9B*, cg00278366, p = 2.5E-5: *C8orf46*, cg15444185, p = 0.049, fig 4D, full results for all 59 CpGs in table S6). As observed in the MZ cohort, hierarchical clustering and PCA of this probe list in the case-control cohort does not indicate any power to discriminate between ALS and control samples (fig. 4E, F).

Differentially methylated probes identified within discordant MZ twin/triplet sets implicates new genes and existing ALS genes

Given the clinical heterogeneity in our twin/triplet cohort, within-twin-set differential methylation was also investigated. Using a threshold of a difference in β-methylation ≥ 0.25 between co-twins or the affected triplet and the mean of the unaffected triplets, we identified 0 DMPs in female SALS twins, 6 in *C9orf72* twins, 58 in *SOD1* triplets, 2,689 in male SALS, and 29 in control twins (fig. S3A-E). Up to 11 probes were annotated per gene in the

male SALS twin list of DMPs, for a total of 1829 genes identified. The 506 genes to which multiple male SALS twin probes annotate are given in table S7, which includes two genes previously associated with ALS, *DPP6* (Dipeptidyl Peptidase Like 6) and *RAMP3* (Receptor Activity Modifying Protein 3) (fig. 5A). No other discordant twin/triplet set had multiple probes annotated to the same gene. Across all discordant twin/triplet sets, 2 probes (fig. 5B) and 13 genes (*BDKRB2, CHRD, DYSF, HOXD11, IRX4, ISL1, JOSD1, mir_544, NKX2-5, NXN, OTX1, POU4F2, RFX4*, fig. 5C) were identified in multiple sets. None of these probes or genes were also identified in the control twin set. Each of the male SALS twins' DMPs, *C9orf72* twins' DMPs and *SOD1* triplets' DMPs showed minimal overlap with the control twins DMPs (5, 1, and 1 DMPs, respectively, fig. 5B, table S8). Similarly, minimal overlapping genes-annotated-to-DMPs were identified between the control and discordant twins/triplets, with 9, 1, and 1 genes respectively (fig. 5C, table S8).

Transcriptome-wide analysis of disease discordant MZ siblings

Differentially expressed genes within male SALS twins implicates immune function and cell signalling functional pathways in sporadic ALS

Using limma voom to detect genes differentially expressed between male SALS twins while controlling for repeated sampling, we identified 4179 genes as significant following FDR correction ($p < 0.05$). Of these, 750 genes also had a fold change of 1.5 or greater (fig 6A, top genes shown in fig 6C, full list in table S9). Notably, *CCNF* and *CCS*, both known ALS genes, were identified as significantly downregulated in the ALS twin compared to their unaffected co-twin (*CCNF*: logFC = 0.70, t = 3.99, FDR = 0.027; *CCS*: logFC = 0.70, t = 6.42, FDR = 0.008, figure 6B). Gene Ontology (GO) analysis of these 750 genes identified

74 terms significantly enriched in this list. Over-representation of genes was seen in 25 terms associated with immune function and cell signalling, while there was an under-representation of genes associated with 45 terms, largely related to transcription and protein modification (fig. 7, table S10).

Validation of twin differentially expressed genes in a case-control cohort

Within the validation data set of SALS and controls, 379 of the 750 genes identified in the male SALS twins were present. When analysed with limma while controlling for sex, 213 of the 379 genes were differentially expressed between cases and controls, yet none also showed a minimum fold change of 1.5 (top 8 of 213 genes shown in fig. 8A, 213 genes in table S11). *CCNF* was not present in the case-control data set, while *CCS* was not validated (log FC = 0.13, t = 1.99, FDR = 0.075). Hierarchical clustering and PCA of the 379 genes did not identify clusters representing disease status (fig. 8B, C).

Integration of genome-wide methylation and transcriptome data sets

To increase the likelihood of detecting biologically meaningful disease-related alterations, RNA-Seq and Infinium HumanMethylation450K data sets were combined for the male SALS twins.

Shared overlap between RNA-Seq and Infinium HumanMethylation450K data sets

Of 506 genes having at least one differentially methylated CpG probe annotated to them in the male SALS twins, 123 are also identified in the entire post-processing RNA-Seq data set of 13718 genes. Conversely, of the 750 genes present in our top DEG list, 642 also have at

least one CpG probe mapped to the same gene in the full post-processing 450K set of 24073 genes.

Integration

The 123 genes identified as differentially methylated in the male SALS twins and present in the full gene expression data set, and the 642 genes found to be differentially expressed in the same twins and to have one or more CpGs in the full 450K methylation dataset, were compared. Of these, 12 genes (*C11orf49, CD8A, COL7A1, EOMES, GATA6, GZMM, HOXA4, KANK3, OLIG2, QPRT, SMPD3, SNED1*) were present in both gene lists (fig. 9A-C).

Discussion

Using a longitudinal cohort of MZ twins and triplets that are discordant for ALS, have conducted both a targeted and genome-wide DNA methylation study in conjunction with a matched sample transcriptomic study. Our cohort is representative of the clinical heterogeneity (age of disease onset, disease duration) frequently observed in ALS cohorts. We have shown that DNA methylation age is the most consistently altered epigenetic signature in ALS. In addition, we observed a higher frequency of unique peripheral blood methylation changes within twin/triplet sets compared to shared methylation changes across twin/triplet sets. However, combined analysis of peripheral blood methylation and transcription detected ALS-relevant changes. These data suggest that the epigenetic and transcriptomic landscape of ALS may be highly complex with numerous small perturbations and various pathways, only some of which are common, contributing to disease.

Epigenetic age was significantly associated with disease in an age-dependent manner, such that affected twins/triplets have an older DNA methylation age than their unaffected co-twins/-triplets while no such effect was observed in young discordant twins. A clear and consistent difference was apparent between the oldest twins in the study, and to a lesser extent, within both middle-aged twin/triplet sets. This pattern of increased methylation age in ALS affected twins is consistent with previous studies [27, 28]. Increased DNA methylation age has been linked to increased mortality [32] and age has been shown to be a major risk and prognostic factor for ALS [33]. Our results also reflect a contribution of ageing to disease risk. Methylation age has also been previously linked to age of onset in ALS patients with a *C9orf72* repeat expansion [34], while we observed a similar phenomenon in our sporadic ALS twin sets, with a much greater between-co-twin difference in DNA methylation age in our late onset twin set compared to our early onset twin set. Further investigation in extended ALS cohorts, specifically mutation-known FALS and SALS would be worthwhile to confirm the contribution of increased DNA methylation age to ALS.

When assessing genome-wide DNA methylation using a magnitude and statistical ranking method, we identified 59 probes differentially methylated in all ALS twins/triplets compared to their unaffected co-twin/-triplets. These 59 probes were selected from high CpG density regions of the genome, therefore considered biologically relevant as they are more likely to affect gene expression. Annotation of the probes to the closest gene transcription site and subsequent gene ontology analysis implicated developmental processes. However, clustering of these 59 DMPs were unable to discriminate between affected and healthy twins, or sporadic cases and controls. Yet, two of these probes were validated as significantly differentially methylated in the case-control analysis. *C8orf46*'s *Xenopus* homolog *vexin* is involved in neurogenesis and highly expressed in the brain [35], while *RAD9B* responds to

DNA damage by moving to the nucleus and contributes to control of the cell cycle [36]. There is a growing body of evidence that DNA damage response is a significant factor in ALS [37].

We conducted within-twin/triplet-set comparisons to show the *SOD1* triplet set, *C9orf72* twin set and the male SALS twin set each have a moderate number of probes with large differences in methylation (6, 58, and 2689 probes respectively with $|\Delta\beta| \geq 0.25$). In contrast, the female SALS twin set showed highly consistent methylation across all >386,000 probes (max $|\Delta\beta| = 0.11$). It is noteworthy that our four twin sets represent two distinct genetic forms of disease (*SOD1* and *C9orf72*), along with two cases at extreme ends of the clinical spectrum of sporadic ALS, suggesting again that there may be various epigenetic pathways impacting the phenotype. Some proportion of the observed differences unique to a twin set may result from epigenetic drift [38], especially as the greatest number of unique differentially methylated probes was identified in the oldest twin set and the least in the youngest twin set. It is therefore likely that disease, as well as age, is contributing to the differential methylation observed. Nevertheless, we identified multiple differentially methylated probes annotated to two genes previously associated with ALS, *DPP6* [39] and *RAMP3* [40] in our oldest twin set, the male SALS twins. *DPP6* was the first gene to be associated with sporadic ALS [41]. It has roles regulating dendritic excitability [38], with membrane hyperexcitability observed in ALS [42, 43]. It has also been associated with multiple sclerosis [44] and spinal muscular atrophy [45], and as such is worthy of further investigation in broader ALS.

Analysis of gene expression in a subset of our disease discordant MZ cohort, the male SALS twins, found 750 differentially expressed genes. 379 of these were tested in our validation

sporadic case-control cohort, and 213 were confirmed. Gene Ontology analysis implicated primarily upregulation of the immune system, which has been previously identified as dysregulated in ALS [16, 18, 19, 29]. Interestingly, *CCNF* and *CCS* were downregulated in the ALS-affected twin. *CCNF* has been identified as a causal ALS and FTD gene in several international cohorts [46]. While transient overexpression has been shown to have deleterious effects in *CCNF* zebrafish models [47], this is the first report of altered *CCNF* mRNA expression in ALS. *CCS* has been previously linked to *SOD1* in its implication in ALS [48]. Little is known about the effects of altered expression of *CCS* in ALS, but its overexpression in the G93A-SOD1 ALS mouse model has been linked to accelerated neurological deficits and worsened mitochondrial pathology [49]. It is interesting that we observed lower expression in the ALS twin than their unaffected co-twin, given that overexpression has been linked to disease in both genes. It was an unfortunate limitation of this study that neither gene featured in our post-processing HumanMethylation450K dataset, and that so few of the genes identified had data available in our case-control data set. As such, it would be worthwhile to further investigate disease-dependent expression of the remaining 371 genes.

Comparison of transcriptional and DNA methylation changes in ALS-discordant twin/triplet set(s) indicated that despite many genes being present in only one data set, there remained overlap between the two. Of the 750 differentially expressed genes identified in the male SALS twins, 642 had any methylation data available, while of the 506 genes to which multiple of the 1366 differentially methylated probes annotated, only 123 were also represented in our gene expression data. When we compared these 642 expression-derived genes and 123 methylation-derived genes, we identified twelve genes: *C11orf49, CD8A, COL7A1, EOMES, GATA6, GZMM, HOXA4, KANK3, OLIG2, QPRT, SMPD3, SNED1*. Notably, ALS genes identified as DMPs in the male SALS twin set, *RAMP3* and *DPP6*, were

not present in the post-processing male SALS twins RNA-Seq data set. *C8orf46* and *RAD9B* were identified across all twin sets to have a single probe differentially methylated, which was validated in our sporadic case-control cohort, however, neither gene was present in our RNA-Seq data set. While *CCNF* and *CCS* were differentially expressed in the male SALS twins, neither gene was present in the methylation dataset.

While none of the twelve genes have previously been directly linked to ALS, some indirect links exist. *COL7A1*, as part of the collagen gene family, is related to *COL6A1*, which has been linked to neurodegeneration through impaired autophagy and induction of apoptosis [50]. Additionally, collagen has also been identified as a significant gene ontology term in analysis of DNA methylation in sporadic ALS [13]. *GZMM*, granzyme M, is 1 of 4 gene products from the granzyme family. Granzymes A and B are elevated in ALS serum, with granzyme B correlated to ALS severity [51]. Granzyme B has been further implicated in inducing apoptosis in human ALS motor neurons [52]. *SMPD3*, neutral Sphingomyelinase II, is associated with apoptosis and cell cycle regulation, which have been previously linked to ALS [53, 54]. *KANK3* has been suggested as a possible gene contributing to an ALS-linked region on chromosome 17 [55]. *QPRT* is involved in the kynurenine pathway, which has been implicated in ALS [56].

These twelve genes, identified when combining DNA methylation and gene expression data, may thus contribute to disease, and warrant further investigation.

Assessment of global methylation and blood cell composition showed no difference between ALS and healthy co-twins. Although a lack of global changes in methylation is consistent with five other sets of ALS-discordant twins [27], not all studies agree [12, 16, 17]. It is also interesting that blood cell composition, as determined from whole blood methylation, was not found to vary between affected and unaffected twins, given that upregulation of the immune

system and changes in white blood cell populations have previously been demonstrated in ALS [57, 58]. This lack of effect in white blood cell estimates may be partly attributable to shared genetic background [59, 60], although a prior study reported differing methylation-derived cell proportion estimates in one ALS-discordant twin pair [29].

High-density quantitative targeted analysis of the *C9orf72* and *SOD1* gene-associated CpG islands and gene promotors did not identify any differences in methylation status between ALS-discordant MZ twins/triplets carrying mutations in these genes. The general consistency observed in *SOD1* methylation between carriers of *SOD1* mutations suggests DNA methylation of the *SOD1* promoter itself is not likely to be a major mechanism contributing to differences in penetrance in *SOD1*-linked ALS, in line with previous reports[12].

Methylation of *C9orf72* was low in a twin set carrying the *C9orf72* repeat expansion. Methylation of the *C9orf72* promoter and/or the repeat expansion has been reported in the brain and blood of repeat expansion carriers [8, 11, 61-65], in some cases with similar low levels of methylation as that observed here. Interestingly, neither of the two prior *C9orf72* twin studies, one ALS concordant and one discordant, detected methylation of *C9orf72* [26, 66], suggesting that *C9orf72* methylation is just one part of the epigenetic story in ALS.

Conclusion

In conclusion, our disease-discordant twin study, utilising longitudinal samples throughout disease progression, demonstrated significant association of DNA methylation age with disease in an age dependent manner. We have also identified an important set of DMPs and DEGs, and associated functional pathways, that may be involved in either ALS pathogenesis

or protection from disease. These genes and pathways offer potential targets for future therapeutic treatment for ALS patients.

Methods

Participants

The cohort of 1806 total participants used in this study is summarised below. Samples from ALS patients, family members, and unrelated controls were obtained from the Macquarie University Neurodegenerative Diseases Biobank, Molecular Medicine Laboratory at Concord Hospital, and the Australian MND DNA bank. All individuals were recruited under informed written consent as approved by the human research ethics committees of Macquarie University and Sydney South West Area Health Service. Most participants were of European descent and patients were clinically diagnosed with definite or probable ALS based on El Escorial criteria [67]. Genomic DNA was extracted from peripheral blood using standard protocols. RNA was extracted from peripheral blood with the QIAsymphony PAXgene blood RNA kit (Qiagen, Hilden, Germany).

Twin/triplet cohort

Three ALS discordant monozygotic twin pairs, one ALS discordant MZ triplet set and two control MZ twin pairs were included in this study (fig. 1 and table 1). Monozygosity for each twin/triplet set was confirmed using STR fragment analysis and/or SNP microarrays. Longitudinal samples were available from two twin sets (male and female sporadic ALS (SALS) twin sets in fig. 1). The four discordant twin/triplet sets had previously undergone

mutation analysis for known ALS genes and whole genome analysis for novel and/or rare *de novo* variants.

Data processing and validation cohorts

Additional samples were used in this study for data processing (Illumina HumanMethylation 450K, EpiTYPER methylation assays, and RNA-Seq) and for validation of significant findings (Illumina HumanMethylation 450K assay and RNA-Seq).

For quality control and processing of the EpiTYPER data, 279 samples with *C9orf72* EpiTYPER data (158 familial ALS/FTD samples, 56 asymptomatic samples (individuals harbouring a causal gene mutation but currently unaffected), and 65 control samples), and 261 samples with *SOD1* EpiTYPER data (123 familial ALS, 65 asymptomatic, and 73 control samples) were used.

For the Infinium HumanMethylation 450K BeadChip, 1658 samples were used in data processing and normalisation. This comprised 889 individuals with sporadic or familial ALS, 92 asymptomatic and 668 controls. The familial ALS and asymptomatic cases largely overlap with the EpiTYPER cohort. The validation subset comprised 650 sporadic ALS individuals and 539 unrelated controls.

One hundred and ninety samples were used for data processing and normalisation of the RNA-Seq data, comprising 114 individuals with ALS (99 sporadic ALS, 15 familial ALS) and 76 unrelated controls. The validation subset comprised of 96 sporadic cases and 96 controls. The majority of the 96 validation sporadic ALS cases were also present in the HumanMethylation 450K BeadChip SALS validation cohort.

Demographic characteristics between cases and controls in validation cohorts were assessed with t-tests for age and $\chi 2$ tests for sex.

Methylation assays and data processing

All quality control and data processing steps were carried out in R v 3.4.4 [68].

EpiTYPER assay

Custom EpiTYPER assays (Sequenom, San Diego, USA) were used to quantify CpG methylation of 56 and 39 CpG units respectively of the two gene-associated CpG islands for *C9orf72* and the gene-associated CpG island upstream of *SOD1*. EpiTYPER uses base-specific cleavage of bisulphite-converted DNA and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MADL-TOF MS) to quantify DNA methylation [69]. Primers for overlapping amplicons were designed with Sequenom's EpiDesigner software to target the CpG island regions, and therefore the promoter regions, as shown in fig. S1. Primer and assay details are available in table S1. Samples were assayed in one or two batches, and either in duplicate or as singletons (table S1). Sample processing was performed by Agena Bioscience (Brisbane, Queensland, Australia). As each gene assay was run across several plates of samples, the highly methylated DNA control was used to calculate the between-plate coefficient of variation, determined to be 4.9% and 2.3% for *SOD1* and *C9orf72* plates, respectively. CpG methylation was quantified as the percentage of methylated cytosines for each CpG unit, where CpG units consist of one or more CpG sites. For units with multiple CpG sites, methylation percentages were normalised by averaging across the number of sites.

EpiTYPER data processing

EpiTYPER data processing was adapted from a previously established method [70]. Twin samples were processed together with the full familial cohorts to leverage the increased

sample size. In brief, CpG units that failed to meet assay reliability standards were discarded and samples in duplicate were averaged for the remaining CpG units. Given the relatively small number of CpG units remaining after removal of those determined to be unreliable and the relatively high failure rate of samples and units, a two-step sample / unit filtering process was used. First, failed samples, with ≥ 90% of CpG unit readings missing, were removed, followed by CpG units which were missing data for ≥ 90% of samples. Second, samples with a low detection success (missing data for ≥ 15% of units) were removed, and the same threshold applied to remove CpG units with low detection success (units missing data for ≥ 15% of samples). Finally, any remaining missing values were imputed with the mean for that unit. Following data processing and filtering, 28 of 56 and 23 of 39 CpG units (for *C9orf72* and *SOD1*, respectively) remained for analysis.

Infinium Human Methylation 450K v1.2 BeadChip array

Genome-wide methylation was investigated using the Infinium HumanMethylation 450K v1.2 BeadChip (Illumina, San Diego, USA). This microarray provides qualitative methylation values for approximately 480,000 CpG sites distributed throughout the genome. Bisulphite-converted DNA was hybridised to the Infinium HumanMethylation 450K BeadChip. Fluorescence imaging of the BeadChip using an Illumina HiScan SQ scanner successfully generated raw Intensity Data files (.idat) for all samples.

450K data processing

Data processing of the .idat files was adapted from the method presented by [71]. Twin samples were processed together with the full cohort to leverage larger sample sizes. All default settings were used except where otherwise specified. In brief, samples with less than 99% of CpGs detected were removed. shinyMethyl (v. 1.12.0, [72]) was used to visually

identify possible outliers, with confirmation of sex queries using RnBeads (v. 1.0.0., [73]). Samples with any possibility of incorrect identification were removed. Data were normalised with the dasen function from wateRmelon (v. 1.20.3, [74]). Probes that had failed to be detected (threshold $p > 0.05$) with the minfi (v. 1.22.1, [75]) function detectionP were removed (n=10,270 probes). Normalised data were submitted to Horvath's DNAm age calculator [30]. Samples that did not strongly correlate ($r < 0.85$) with the DNAm age results gold standard were removed. Leveraging technical replicate/duplicate samples (n=30), both 1) in the form of multiple blood collections at the same time and resulting independent DNA extractions (technical replicates) and 2) multiple aliquots of single DNA extractions (duplicates), a custom filtering step was included to identify and remove highly variable probes. Any probe identified to have multiple pairs of technical replicate or duplicate samples with differences greater than three standard deviations from the probe's mean difference was discarded (n=38,697). Of the remaining probes, any known to cross hybridise, be located on sex chromosomes, or bind to SNPs, were removed (n=50,362) [76].

Following raw data processing, quantitative CpG methylation values for 1,215 samples (including 34 twin samples from table 1 and 1,179 case/control validation cohort samples outlined in table S2) and 386,183 probes remained for analysis and validation. Comparison of the case-control validation cohort (table S2) showed that sex ($\chi^2_{(1,\, n\, =\, 1179)} = 33.8$, $p < 0.01$) and age ($t_{1174} = 4.20$, $p < 0.01$) were significantly different between ALS cases and controls.

Analysis of methylation data

All statistical analyses were carried out in R v. 3.4.4 [68].

Gene-specific targeted methylation analysis of *SOD1* and *C9orf72* in the FALS twin/triplet sets

Methylation of *SOD1* or *C9orf72*, as quantified by both EpiTYPER and 450K assays, was visualised in the relevant monozygotic disease discordant twin/triplets. Four and five 450K CpGs were available in the post processing data set in the targeted region of *C9orf72* (cg05990720, cg11613875, cg14363787, cg23074747) and *SOD1* (cg16086310, cg17253939, cg18126791, cg19948014, cg26893544), respectively. Since only one twin set and one triplet set are available in our cohort for each respective variant, results are descriptive only.

Observed differences in DNA methylation age, blood cell composition and global methylation within a twin/triplet set

DNA methylation age was determined from 450K methylation data using the method of Horvath [30].

Blood cell proportions in whole blood derived methylation was estimated from 450K methylation data with the minfi implementation of Houseman et al.'s [31] algorithm.

Global methylation levels were determined as the mean methylation estimate across all post-processing 450K CpG sites per sample. CpG sites were also divided into one of four categories based on HIL CpG classes [77] and the mean methylation for each was calculated.

Methylome-wide analysis in MZ sets to identify differentially methylated probes

The list of differentially methylated probes (DMP) across all MZ sets was identified using a ranked magnitude-significance method [78]. In brief, statistical significance per CpG site was determined using a paired t-test on methylation M-values, using the per-patient mean of longitudinal samples and unaffected triplets. The magnitude of the difference in methylation

was calculated as the mean difference in β-methylation between co-twins. Both methods were used to rank all CpGs, and a final ranked list was determined from the mean of these two ranking methods. Top DMPs were the subset of all CpG probes that met the following two criteria, 1) they were in high CpG density regions of the genome and 2) the ranked list of high-density probes was truncated immediately prior to the first probe to show a difference in the direction of change across the four discordant MZ sets. The ability of these probes to discriminate between ALS and healthy individuals was assessed by hierarchical clustering of all twin/triplet sets.

Within-twin/triplet set DMPs were also identified. A CpG probe was considered to be differentially methylated within a twin/triplet set where there was an absolute difference in β-methylation ≥ 0.25 between the affected twin and their unaffected co-twin/triplet.

Validation of identified twin DMPs in a sporadic ALS cohort

Twin/triplet DMPs were validated in the larger sporadic case-control cohort. Differences between cases and controls for each of the identified probes were analysed, along with the ability of the DMP list to cluster cases and controls separately.

Gene expression

RNA sequencing

Raw sequencing reads in fastq format were generated for male SALS twins (based on longitudinal sample availability) and the sporadic case-control validation cohort as outlined in table S3.

RNA-Seq data processing

The quality of raw sequencing reads was evaluated using fastQC (v 0.11.7 [79]) for both datasets. Trimming and alignment was performed as outlined in table 1 using either Trimmomatic (v. 0.36 [80]) or Cutadapt (version 1.8.1, [81]) and HISAT2 (v2.0.5 82]). All subsequent data processing and analysis was completed in R (v. 3.4.4), using BioConductor packages edgeR v. 3.18.1 [83] and limma v. 3.32.10 [84]. A standard edgeR TMM normalisation and filtering pipeline was used in data processing, with only those genes where expression was greater than 0.3 counts per million in a minimum of 3 samples (male SALS twins) or 2 counts per million in a minimum of 75 samples (case-control cohort) retained for analysis, which is equivalent to approximately 12-15 raw counts in the smallest library size for each dataset. For the male SALS twins RNA-Seq data, of the 27,685 human genes present in the per-gene read counts generated by HTSeq [85], 13,718 genes remained following raw data processing using edgeR [83]. Whereas in the case-control cohort, of the 23,368 human genes present in the per-gene read counts generated by HTSeq [85], 7,354 genes remained following raw data processing using edgeR [83].

MDS (multi-dimensional scaling) indicated the presence of three outliers in the case-control cohort, 1 control and 2 SALS samples. All three were removed and final clinical details for the cohort can be found in table S4. Comparison of the RNA-Seq case-control validation cohort (table S4) showed that there were no significant differences in age ($t_{157.9}$=1.74, p=0.08) between the ALS cases and controls, but a difference was observed in sex between cases and controls ($\chi^2_{(1,\ n\ =\ 165)}$=6.5, p=0.01).

Differentially expressed genes in MZ twins

To identify differentially expressed genes (DEGs) using the paired longitudinal RNA-Seq samples from the ALS-discordant male SALS twins (table 2), read count data was analysed using limma [84], including model terms for longitudinal sample collection and disease

status. Voom [86] transformation was applied prior to modelling. Multiple testing correction using the BH-FDR method [87] was applied to the full list of post-processing genes.

Validation of twin DEGs in a sporadic cohort

Genes identified in twin analyses were investigated for an effect of disease in the full case-control cohort with limma, including sex as a covariate. Data were voom transformed, given the highly variable library sizes. Multiple testing corrections using the BH-FDR method was applied only on the subset of genes identified as differentially expressed in twins. Hierarchical clustering of the expression of these DEGs in the case-control cohort was assessed.

Combined methylation and expression analysis

Intersect of top CpGs and genes

To identify genes most likely to be altered in disease, results from independent analysis of genome-wide methylation and expression data sets were integrated. Longitudinal RNA-Seq data is only available for one male SALS twin set, therefore we first identified the overlap between top DEGs and the genes annotated to the most differentially methylated probes within that twin set. We extended this analysis by overlapping the same list of DEGs with the genes annotated to the top differentially methylated probes across all combined twin sets.

Gene Ontology analysis

Gene Ontology enrichment analysis [88, 89] for biological processes was applied to the genes identified as differentially expressed in male SALS twins. The gene list was analysed with PANTHER overrepresentation tests (GO Ontology database release 2018-08-09, [90]). Enrichment was tested relative to all genes detected in the appropriate post-processing data set. Fisher's exact test with FDR correction was used.

Statistics

All analyses were carried out in R (v. 3.4.4) [68]. Linear mixed effects models were used to analyse DNAm age, blood cell type proportions and global mean M-methylation. Modelling was carried out using the lmer function in the package lme4 (v. 1.1.14, [91]) for DNAm age and mean methylation, while a mixed effects beta regression for cell type proportions was applied with the glmmTMB function from the glmmTMB package (v. 0.2.2.0, [92]). Blood cell type proportions were increased by 0.001 to all estimates to avoid taking the log of zero. All mixed models assessed the effect of disease status while controlling for age at sample collection and sex. When analysing DNAm age, the interaction of disease and age at collection was also tested. Random effects were introduced for repeated sampling within co-twins, and a random intercept per twin/triplet set. When modelling cell types, due to convergence issues, the random slope for repeated sampling was dropped, leaving random intercepts for each co-twin and twin set. Likelihood ratio tests were used to determine significance of model terms.

Linear models were used for case-control validation of probes identified in the MZ cohort, with the same fixed effect terms of age at sample collection and sex as described for mixed models.

Hierarchical clustering utilised the package cluster (v. 2.0.6), with Manhattan distance and ward clustering methods for 450K data [93], and Spearman correlation distance and average linkage clustering for log-transformed RNA-Seq count data [94].

Where appropriate, technical replicates are shown as means with error bars indicating standard deviation (unless otherwise stated).

Declarations

Ethics approval and consent to participate

All individuals were recruited under informed written consent as approved by the human research ethics committees of Macquarie University (5201600387) and Sydney Local Health District.

Consent for publication

Not applicable

Availability of data and material

The datasets generated and/or analysed during the current study are not publicly available since our ethics permission does not cover sharing of data to third parties but are available from the corresponding author [KW] on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was funded by the Motor Neurone Disease Research Institute of Australia (grant to KLW), National Health and Medical Research Council of Australia (grant 1083187 to IPB, fellowship 1092023 to KLW) and Macquarie University (grant to KLW).

The funding bodies did not play a role in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Authors' contributions

IST and KLW conceived and designed the study with input from SJC, NW and IPB. IST, EM, BB, TJP, NAT, KYZ, QZ, Z-HZ, DB and KLW performed the experiments/data analysis and interpretation. DBR and GAN collected clinical information and samples. IST and KLW wrote the manuscript with input from IPB. All authors read and approved the final manuscript.

Acknowledgements

We thank Carolyn Cecere and Lorel Adams for their assistance in compiling family information, Elisa Cachia and Dr Sarah Furlong for providing patient materials, clinical and technical assistance, and Janette Edson for technical assistance.

Legends

Fig 1. ALS-discordant twin/triplet set pedigrees

Pedigrees for four sets of ALS-discordant twins/triplets, with gene mutations indicated. Circles represent females and squares represent males. Filled shapes indicate ALS, open shapes with a dot indicate mutation carriers and open shapes are unaffected non-carriers. Horizontal lines between twins/triplets indicate monozygosity.

Fig 2. Neither *C9orf72* nor *SOD1* are differentially methylated between mutation-positive ALS-discordant twins/triplets

(**A**) Methylation of the CpG island spanning the promoter region and exon 1 of *SOD1* does not show differential methylation between an ALS-affected triplet and unaffected co-triplets, concordant for *SOD1* p.I114T. Methylation status was determined using both EpiTYPER (bottom) and 450K (middle) assays. The relative location of targeted CpG islands (CGI) and exon 1 are indicated for both genes (**A** and **B**, top). (**B**) Methylation of the *C9orf72* promoter region / expansion flanking CpG islands are not differentially methylated between ALS-discordant co-twins that carry the *C9orf72* hexanucleotide repeat expansion in either EpiTYPER (bottom) or 450K data sets (middle). Transcript variants (T1, T2, and T3) and the position of the repeat expansion (black diamond) relative to exon 1 are shown for *C9orf72* (**B**, top).

Fig. 3. DNA methylation age, but neither global mean methylation nor cell composition, varies between ALS-discordant twins/triplets

(**A**) DNA methylation age was more discrepant between ALS discordant twins with increasing chronological age (p = 1.3E-5), with greater DNAm aging in affected twins/triplets than their unaffected co-twin/-triplets, when controlling for age and sex. (**B**) Mean methylation across 386183 CpG sites found no significant difference in global methylation between ALS-affected and unaffected co-twins/-triplets when controlling for age and sex (p = 0.08). Proportions of six white blood cell types over time were estimated for ALS-affected and unaffected twins/triplets (**C**). Proportions for each cell type were not significantly associated with disease status of twin/triplet samples when controlling for age and sex (CD4+ T cells, p = 0.77; CD8+ T cells, p = 0.24; Monocytes, p = 0.60; B cells, p = 0.21; Natural killer cells, p = 0.52; granulocytes, p = 0.63).

Fig 4. Top DMPs identified and validated in MZ and case-control cohorts don't cluster by disease

(**A**) Of 59 DMPs found across all discordant twin sets, these 9 were top ranked for the combination of statistical significantly differences between affected and unaffected co-twins/-triplets and the magnitude of differences across twin/triplet sets. Per twin/triplet set differences are shown, with the ALS-affected sibling as the reference. Gene annotation and CpG name are indicated as *gene::cpg.* Control twins are shown in grey, with the same directionality as the discordant twins to facilitate comparison of the magnitude of the difference in methylation. Bar colour indicates hypomethylation of the ALS-affected twin (orange) or hypermethylation of the affected twin/triplet (blue) relative to their unaffected co-twin/-triplets. (**B**) The 59 DMPs identified across discordant twin/triplet sets were used to cluster the samples in the MZ cohort. Overall, methylation was similar across samples for most DMPs, and samples did not cluster by disease, nor perfectly by twin/triplet set. (**C**) Principal Components Analysis (PCA) across discordant twin/triplet sets using the same 59

DMPs also showed that samples did not cluster by disease, but approximately by individual for those where longitudinal samples were available. (**D**) Of the 59 DMPs identified across discordant twin/triplet sets, two were significantly different between cases and controls in a large cohort (n SALS = 646, n controls = 533). Both cg15444185, annotated to *C8orf46*, and cg00278366, annotated to *RAD9B*, were hypomethylated in SALS samples (cg15444185, β = -0.06, adjusted p = 0.049; cg00278366, β= -0.0771, adjusted p = 2.5E-5) when controlling for age and sex. (**E**) The top 59 DMPs identified across all discordant twin set do not cluster by disease status in a sporadic case control cohort. (**F**) PCA also demonstrates that these top 59 twin DMPs do not cluster by disease status in a sporadic case control cohort.

Fig 5. Most differentially methylated probes (DMPs) per twin set were unique to one twin set including known ALS genes

DMPs within a twin/triplet set were those with a difference in β-methylation ≥ 0.25. (**A**) Within the male SALS twin set, two probes were identified which annotated to *DPP6*, and two additional probes annotated to *RAMP3*. Multiple data points per person at each probe indicate longitudinal sampling. For collection times with duplicate samples per person, points represent the mean at that time, with the standard deviation indicated with a line. (**B-C**) Generally, DMPs were unique to a twin set, while no differences in methylation (> 0.25) were detected in the female SALS twins. (**B**) The number of DMPs within a twin set varied from < 10 in *C9orf72* twins to > 2500 male SALS twins (fig. S3B,D). Only two of these DMPs were found in multiple discordant twin sets. Each of the male SALS twins, *SOD1* triplets and *C9orf72* twins showed overlap with the control twins. (**C**) Within each of the three discordant twin sets and the control twin set DMP lists, multiple probes annotated to the same gene. When comparing these genes rather than individual probes, more shared genes

were identified between discordant sets, with 13 genes containing a probe considered differentially methylated in multiple discordant twin/triplet sets.

Fig 6. Genes that showed consistent longitudinal differential expression in SALS twins included known ALS genes

(**A**) Seven hundred and fifty genes were identified as differentially expressed with a minimum fold change of 1.5 (vertical lines) and significant FDR-corrected p-value (horizontal line) in the male SALS twins. (**B**) Expression of two previously reported ALS genes, *CCNF* and *CCS*, identified as differentially expressed in male SALS twins. Gene expression is shown for all three collections in each twin. (**C**) Expression of the top 8 genes (as ranked by limma) are shown for all three collections of the male SALS twins.

Fig 7. Significantly enriched Gene Ontology (GO) terms implicate enrichment of immune function in the ALS co-twin.

GO analysis of the 750 longitudinally differentially expressed genes from the male SALS twins identified 74 significantly enriched biological processes or pathways, shown on the y-axis of the graph. Adjusted p-value (using FDR method) is indicated by the height of the columns on the graph (x-axis). Log2 fold enrichment (logFoldEnrichment) of GO biological process is indicated by depth of colour, and direction of gene representation (red = over-representation in affected co-twin, blue = over-representation in affected co-twin). Results demonstrate over-representation of genes associated with immune function and cell signalling, and under-representation of genes largely related to transcription and protein modification.

Fig 8. DEGs identified in male SALS twin are validated in a case-control cohort

Of the 750 DEGs identified in the male SALS twins, only 379 genes were present in the sporadic case-control cohort. (**A**) Two hundred and thirteen of these were validated as differentially expressed between SALS and controls when controlling for sex (table S7) and the top 8 are shown here. (**B**) Hierarchical clustering of the sporadic ALS and control cohort by these 379 genes did not identify disease-based clusters. (**C**) Principal Components Analysis (PCA) of the sporadic ALS and control cohort by these 379 genes also did not identify disease-based clusters.

Fig. 9. Twelve overlapping genes were identified in the male SALS twins DMPs and DEGs. (**A**) While 506 genes were identified as having multiple probes with a difference in β-methylation (≥ 0.25) between the ALS-discordant male SALS twins, only 123 of these genes were present in the matching RNA-Seq data. 642 of 750 genes identified as differentially expressed were present in the matching DNA methylation data. Twelve of these genes were both differentially expressed and differentially methylated (*C11orf49, CD8A, COL7A1, EOMES, GATA6, GZMM, HOXA4, KANK3, OLIG2, QPRT, SMPD3, SNED1*). (**B**) On the HumanMethylation 450K beadchip, multiple CpG probes are annotated to each gene. Methylation of all probes annotated to each of the twelve overlapping genes is highly consistent within each co-twin. Distance from the transcription start site in base pairs is shown on the x axis. Multiple data points per person at each probe indicate longitudinal sampling. Duplicate collections within a time point are shown as the mean with the standard deviation indicated by a line. (**C**) Longitudinal expression of the 12 shared genes, (*C11orf49, CD8A, COL7A1, EOMES, GATA6, GZMM, HOXA4, KANK3, OLIG2, QPRT, SMPD3, SNED1*), is consistently different between ALS discordant male SALS co-twins over time.

Additional files

Additional file 1: (PDF) **Supplementary Figures S1-S3.**

Additional file 2: (PDF) **Supplementary Tables S1-S4.** Table S1: EpiTYPER assay details. Table S2. Post-filtering clinical summary of Infinium HumanMethylation450K case-control validation cohort. Table S3. RNA sequencing and processing summary for longitudinal male SALS twins and case-control cohort. Table S4. Post-filtering clinical summary of RNA-Seq case-control validation cohort.

Additional file 3: (CSV) **Supplementary Table S5:** 59 DMPs identified across all discordant twin sets. Twin set columns show the difference in methylation between co-twins/-triplets as ALS – unaffected. Control twins are shown with the absolute difference in methylation. Delta beta: Mean difference across ALS-discordant twin. Magnitude rank: rank of the probe according to Δβ. Final rank: mean of magnitude and t-test p-value ranks. Gene: gene name corresponding to closest transcription start site to the given probe. TSS distance: distance in base pairs to the closest transcription start site.

Additional file 4: (XLSX) **Supplementary Table S6**. Validation statistics for top 59 twin-DMPs in sporadic case control cohort.Gene: gene name corresponding to closest transcription start site to the given probe. Adjusted R2, F statistic, F numerator DF, F denominator DF, F p-value: model summary statistics. Adjusted p-value: FDR adjusted p values for 'Disease' coefficient across all 59 probes tested.

Additional file 5. **Supplementary Table S7. Genes associated with multiple DMPs identified in the male SALS twins.**

Gene: gene name corresponding to closest transcription start site to the given probe. TSS distance: distance in base pairs to the closest transcription start site. Methylation difference: mean difference in beta methylation between male SALS twins, ALS – unaffected.

Additional file 6. **Supplementary Table S8. Overlap between discordant and control twin sets in genes/probes.** Gene: gene name corresponding to closest transcription start site to the given probe. TSS distance: distance in base pairs to the closest transcription start site. Twin set columns show the difference in methylation between co-twins/-triplets as ALS – unaffected. Control twins are shown with the absolute difference in methylation. Only differences greater than the threshold of 0.25 are shown.

Additional file 7. **Supplementary Table S9. Differentially expressed genes identified in male SALS twins.** Log2 fold change: estimated log2 fold change corresponding to the effect of disease status. Adjusted p-value: BH-FDR adjusted p-values. Log odds: log-odds that the gene is differentially expressed.

Additional file 8. **Supplementary Table S10. Significantly enriched Gene Ontology (GO) terms for 750 differentially expressed genes identified in male SALS twins.** Fold enrichment >1 reflects over-representation of a GO biological process term in the affected co-twin and fold enrichment <1 reflects under-representation of a GO biological process term in the affected co-twin. N genes: number of genes associated with each term. FDR: Multiple testing adjustment of p-values using FDR method.

Additional file 9. **Supplementary Table S11. Statistical summary of the validation of 40 twin-DEGs in a case-control cohort.** Log2 fold change: estimated log2 fold change

corresponding to the effect of disease status. Adjusted p-value: BH-FDR adjusted p-values. Log odds: log-odds that the gene is differentially expressed

References

1. Haverkamp LJ, Appel V, Appel SH. Natural history of amyotrophic lateral sclerosis in a database population. Validation of a scoring system and a model for survival prediction. Brain. 1995;118 (Pt 3):707-19.

2. Nicholson GA. Genes and motor neurone disease. In: M K, editor. The Motor Neurone Disease Handbook. Pyrmont, NSW, Australia: Australasian Medical Publishing Company; 2007. p. 14-25.

3. Iguchi Y, Katsuno M, Ikenaka K, Ishigaki S, Sobue G. Amyotrophic lateral sclerosis: an update on recent genetic insights. J Neurol. 2013;260(11):2917-27.

4. Renton AE, Chio A, Traynor BJ. State of play in amyotrophic lateral sclerosis genetics. Nat Neurosci. 2014;17(1):17-23.

5. McCann EP, Williams KL, Fifita JA, Tarr IS, O'Connor J, Rowe DB, et al. The genotype-phenotype landscape of familial amyotrophic lateral sclerosis in Australia. Clin Genet. 2017;92(3):259-66.

6. Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis. Nat Rev Neurol. 2014;10(11):661-70.
7. Al-Chalabi A, Fang F, Hanby MF, Leigh PN, Shaw CE, Ye W, et al. An estimate of amyotrophic lateral sclerosis heritability using twin data. J Neurol Neurosurg Psychiatry. 2010;81(12):1324-6.
8. Belzil VV, Bauer PO, Gendron TF, Murray ME, Dickson D, Petrucelli L. Characterization of DNA hypermethylation in the cerebellum of c9FTD/ALS patients. Brain Research. 2014;1584:15-21.
9. Xi Z, Rainero I, Rubino E, Pinessi L, Bruni AC, Maletta RG, et al. Hypermethylation of the CpG-island near the C9orf72 G4C2-repeat expansion in FTLD patients. Human Molecular Genetics. 2014;23(21):5630-7.
10. Belzil VV, Bauer PO, Prudencio M, Gendron TF, Stetler CT, Yan IK, et al. Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood. Acta Neuropathologica. 2013;126(6):895-905.
11. Xi Z, Zinman L, Moreno D, Schymick J, Liang Y, Sato C, et al. Hypermethylation of the CpG Island Near the G4C2 Repeat in ALS with a C9orf72 Expansion. The American Journal of Human Genetics. 2013;92(6):981-9.
12. Coppede F, Stoccoro A, Mosca L, Gallo R, Tarlarini C, Lunetta C, et al. Increase in DNA methylation in patients with amyotrophic lateral sclerosis carriers of not fully penetrant SOD1 mutations. Amyotroph Lateral Scler Frontotemporal Degener. 2018;19(1-2):93-101.
13. Morahan JM, Yu B, Trent RJ, Pamphlett R. A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis. 2009;10(5-6):418-29.
14. Oates N, Pamphlett R. An epigenetic analysis of SOD1 and VEGF in ALS. Amyotrophic Lateral Sclerosis. 2007;8(2):83-6.

15. Morello G, Spampinato AG, Cavallaro S. Molecular Taxonomy of Sporadic Amyotrophic Lateral Sclerosis Using Disease-Associated Genes. Frontiers in Neurology. 2017;8(152).

16. Figueroa-Romero C, Hur J, Bender DE, Delaney CE, Cataldo MD, Smith AL, et al. Identification of Epigenetically Altered Genes in Sporadic Amyotrophic Lateral Sclerosis. PLoS ONE. 2012;7(12):e52672.

17. Tremolizzo L, Messina P, Conti E, Sala G, Cecchi M, Airoldi L, et al. Whole-blood global DNA methylation is increased in amyotrophic lateral sclerosis independently of age of onset. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2014;15(1-2):98-105.

18. Andrés-Benito P, Moreno J, Aso E, Povedano M, Ferrer I. Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: implications in frontotemporal lobar degeneration. Aging (Albany NY). 2017;9(3):823-51.

19. Zhao W, Beers DR, Hooten KG, et al. Characterization of gene expression phenotype in amyotrophic lateral sclerosis monocytes. JAMA Neurology. 2017.

20. Ebbert MTW, Ross CA, Pregent LJ, Lank RJ, Zhang C, Katzman RB, et al. Conserved DNA methylation combined with differential frontal cortex and cerebellar expression distinguishes C9orf72-associated and sporadic ALS, and implicates SERPINA1 in disease. Acta Neuropathologica. 2017.

21. van Rheenen W, Diekstra FP, Harschnitz O, Westeneng H-J, van Eijk KR, Saris CGJ, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. PLOS ONE. 2018;13(6):e0198874.

22. Kaut O, Schmitt I, Tost J, Busato F, Liu Y, Hofmann P, et al. Epigenome-wide DNA methylation analysis in siblings and monozygotic twins discordant for sporadic Parkinson's

disease revealed different epigenetic patterns in peripheral blood mononuclear cells. neurogenetics. 2016:1-16.

23. Mastroeni D, McKee A, Grover A, Rogers J, Coleman PD. Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease. PLoS One. 2009;4(8):e6617.

24. Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. 2010;20(2):170-9.

25. Byrne EM, Carrillo-Roa T, Henders AK, Bowdler L, McRae AF, Heath AC, et al. Monozygotic twins affected with major depressive disorder have greater variance in methylation than their unaffected co-twin. Transl Psychiatry. 2013;3:e269.

26. Xi Z, Yunusova Y, van Blitterswijk M, Dib S, Ghani M, Moreno D, et al. Identical twins with the C9orf72 repeat expansion are discordant for ALS. Neurology. 2014;83(16):1476-8.

27. Young PE, Kum Jew S, Buckland ME, Pamphlett R, Suter CM. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. PLOS ONE. 2017;12(8):e0182638.

28. Zhang M, Xi Z, Ghani M, Jia P, Pal M, Werynska K, et al. Genetic and epigenetic study of ALS-discordant identical twins with double mutations in SOD1 and ARHGEF28. J Neurol Neurosurg Psychiatry. 2016;87(11):1268-70.

29. Lam L, Chin L, Halder RC, Sagong B, Famenini S, Sayre J, et al. Epigenetic changes in T-cell and monocyte signatures and production of neurotoxic cytokines in ALS patients. The FASEB Journal. 2016.

30. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013;14(10):1-20.

31. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):86.

32. Christiansen L, Lenart A, Tan Q, Vaupel JW, Aviv A, McGue M, et al. DNA methylation age is associated with mortality in a longitudinal Danish twin study. Aging Cell. 2016;15(1):149-54.

33. Chio A, Logroscino G, Hardiman O, Swingler R, Mitchell D, Beghi E, et al. Prognostic factors in ALS: A critical review. Amyotrophic Lateral Sclerosis. 2009;10(5-6):310-23.

34. Zhang M, Tartaglia MC, Moreno D, Sato C, McKeever P, Weichert A, et al. DNA methylation age-acceleration is associated with disease duration and age at onset in C9orf72 patients. Acta Neuropathologica. 2017:1-9.

35. Moore KB, Logan MA, Aldiri I, Roberts JM, Steele M, Vetter ML. C8orf46 homolog encodes a novel protein Vexin that is required for neurogenesis in Xenopus laevis. Dev Biol. 2018;437(1):27-40.

36. Perez-Castro AJ, Freire R. Rad9B responds to nucleolar stress through ATR and JNK signalling, and delays the G1-S transition. J Cell Sci. 2012;125(Pt 5):1152-64.

37. Coppede F, Migliore L. DNA damage in neurodegenerative diseases. Mutat Res. 2015;776:84-97.

38. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604-9.

39. van Es MA, van Vught PWJ, Blauw HM, Franke L, Saris CGJ, Van Den Bosch L, et al. Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nature Genetics. 2008;40:29.

40. Daoud H, Valdmanis PN, Gros-Louis F, Belzil V, Spiegelman D, Henrion E, et al. Resequencing of 29 candidate genes in patients with familial and sporadic amyotrophic lateral sclerosis. Arch Neurol. 2011;68(5):587-93.

41. Kim J, Nadal MS, Clemens AM, Baron M, Jung SC, Misumi Y, et al. Kv4 accessory protein DPPX (DPP6) is a critical regulator of membrane excitability in hippocampal CA1 pyramidal neurons. J Neurophysiol. 2008;100(4): 1835-1847.

42. Park SB, Kiernan MC, and Vucic S. Axonal Excitability in Amyotrophic Lateral Sclerosis. Neurotherapeutics. 2017;14(1): 78-90.

43. Wainger BJ, Kiskinis E, Mellin C, Wiskow O, Han SSW, Sandoe J, et al. Intrinsic Membrane Hyperexcitability of Amyotrophic Lateral Sclerosis Patient-Derived Motor Neurons. Cell Reports. 2014;7(1): 1-11.

44. Brambilla P, Esposito F, Lindstrom E, Sorosina M, Giacalone G, Clarelli F, et al. Association between DPP6 polymorphism and the risk of progressive multiple sclerosis in Northern and Southern Europeans. Neurosci Lett. 2012;530(2): 155-160.

45. van Es MA, van Vught PWJ, van Kempen G, Blauw HM, Veldink JH, and van den Berg LH, et al. DPP6 is associated with susceptibility to progressive spinal muscular atrophy." Neurology. 2009;72(13): 1184-1185.

46. Williams KL, Topp S, Yang S, Smith B, Fifita JA, Warraich ST, et al. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. Nature Communications. 2016;7:11253.

47. Hogan AL, Don EK, Rayner SL, Lee A, Laird AS, Watchon M, et al. Expression of ALS/FTD-linked mutant CCNF in zebrafish leads to increased cell death in the spinal cord and an aberrant motor phenotype. Human Molecular Genetics. 2017;26(14):2616-26.

48. Beckman JS, Estévez AG, Crow JP, Barbeito L. Superoxide dismutase and the death of motoneurons in ALS. Trends in neurosciences. 2001;24:15-20.

49. Son M, Puttaparthi K, Kawamata H, Rajendran B, Boyer PJ, Manfredi G, et al. Overexpression of CCS in G93A-SOD1 mice leads to accelerated neurological deficits with severe mitochondrial pathology. Proceedings of the National Academy of Sciences. 2007;104(14):6072-7.

50. Cescon M, Chen P, Castagnaro S, Gregorio I, Bonaldo P. Lack of collagen VI promotes neurodegeneration by impairing autophagy and inducing apoptosis during aging. Aging (Albany NY). 2016;8(5):1083-98.

51. Iłżecka J. Granzymes A and B levels in serum of patients with amyotrophic lateral sclerosis. Clinical Biochemistry. 2011;44(8):650-3.

52. Song S. ALS Astrocytes Adopt Natural Killer Properties to Induce Motor Neuron Death: The Ohio State University; 2014.

53. Barbosa LF, Cerqueira FM, Macedo AFA, Garcia CCM, Angeli JPF, Schumacher RI, et al. Increased SOD1 association with chromatin, DNA damage, p53 activation, and apoptosis in a cellular model of SOD1-linked ALS. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2010;1802(5):462-71.

54. Marcuzzo S, Bonanno S, Kapetis D, Barzago C, Cavalcante P, D'Alessandro S, et al. Up-regulation of neural and cell cycle-related microRNAs in brain of amyotrophic lateral sclerosis mice at late disease stage. Molecular Brain. 2015;8(1):5.

55. Sher RB, Heiman-Patterson TD, Blankenhorn EA, Jiang J, Alexander G, Deitch JS, et al. A major QTL on mouse chromosome 17 resulting in lifespan variability in SOD1-G93A transgenic mouse models of amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2014;15(7-8):588-600.

56. Chen Y, Brew BJ, Guillemin GJ. Characterization of the kynurenine pathway in NSC-34 cell line: implications for amyotrophic lateral sclerosis. Journal of Neurochemistry. 2011;118(5):816-25.

57. Mantovani S, Garbelli S, Pasini A, Alimonti D, Perotti C, Melazzini M, et al. Immune system alterations in sporadic amyotrophic lateral sclerosis patients suggest an ongoing neuroinflammatory process. J Neuroimmunol. 2009;210(1-2):73-9.

58. Rentzos M, Evangelopoulos E, Sereti E, Zouvelou V, Marmara S, Alexakis T, et al. Alterations of T cell subsets in ALS: a systemic immune activation? Acta Neurol Scand. 2012;125(4):260-4.

59. Evans DM, Frazer IH, Martin NG. Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. 1999;2(4):250-7.

60. Garner C, Tatu T, Reittie JE, Littlewood T, Darley J, Cervino S, et al. Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood. 2000;95(1):342-6.

61. Gijselinck I, Van Mossevelde S, van der Zee J, Sieben A, Engelborghs S, De Bleecker J, et al. The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. Mol Psychiatry. 2016;21(8):1112-24.

62. Liu EY, Russ J, Wu K, Neal D, Suh E, McNally AG, et al. C9orf72 hypermethylation protects against repeat expansion-associated pathology in ALS/FTD. Acta Neuropathologica. 2014;128(4):525-41.

63. McMillan CT, Russ J, Wood EM, Irwin DJ, Grossman M, McCluskey L, et al. C9orf72 promoter hypermethylation is neuroprotective: Neuroimaging and neuropathologic evidence. Neurology. 2015;84(16):1622-30.

64. Russ J, Liu EY, Wu K, Neal D, Suh E, Irwin DJ, et al. Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. Acta Neuropathologica. 2015;129(1):39-52.

65. Xi Z, Zhang M, Bruni AC, Maletta RG, Colao R, Fratta P, et al. The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. Acta Neuropathologica. 2015;129(5):715-27.

66. Conforti FL, Tortelli R, Morello G, Capozzo R, Barulli MR, Cavallaro S, et al. Clinical features and genetic characterization of two dizygotic twins with C9orf72 expansion. Neurobiology of aging. 2018;69:293 e1- e8.

67. Brooks BR, Miller RG, Swash M, Munsat TL, World Federation of Neurology Research Group on Motor Neuron D. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord. 2000;1(5):293-9.

68. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

69. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proc Natl Acad Sci U S A. 2005;102(44):15785-90.

70. Ho V, Ashbury JE, Taylor S, Vanner S, King WD. Quantification of gene-specific methylation of DNMT3B and MTHFR using sequenom EpiTYPER(R). Data Brief. 2016;6:39-46.

71. Maksimovic J, Phipson B, Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data. F1000Res. 2016;5:1281.

72. Fortin JP, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Res. 2014;3:175.

73. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014;11(11):1138-40.

74. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.

75. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30.

76. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9.

77. Triche, Jr. T (2014). FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes. R package version 2.2.0. (2014)

78. Dempster EL, Pidsley R, Schalkwyk LC, Owens S, Georgiades A, Kane F, et al. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. Hum Mol Genet. 2011;20(24):4786-96.

79. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

80. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

81. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17(1):3.

82. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature methods. 2015;12(4):357.

83. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

84. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-Sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

85. Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.

86. Law CW, Chen Y, Shi W and Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15:R29.

87. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. 1995;57(1):289-300.

88. Consortium TGO. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45(D1):D331-D8.

89. Consortium TGO, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

90. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(D1):D183-D9.

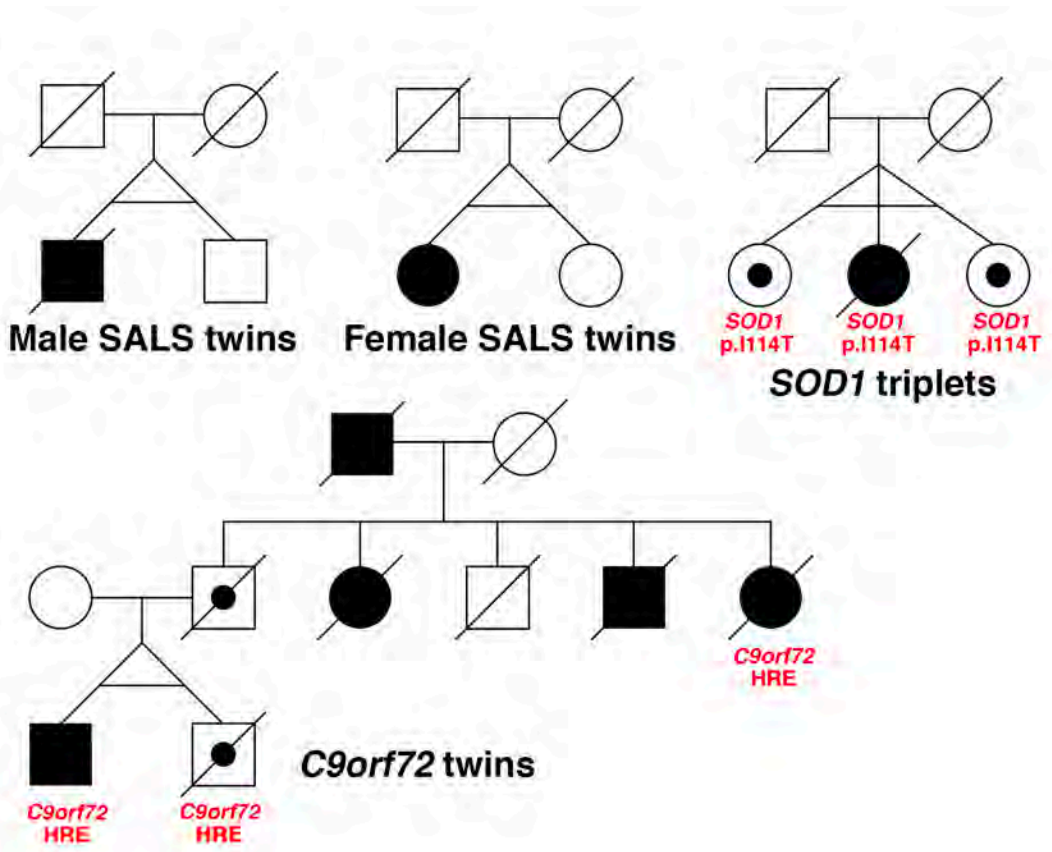
91. Bates D, Machler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015;67(1):1-48.

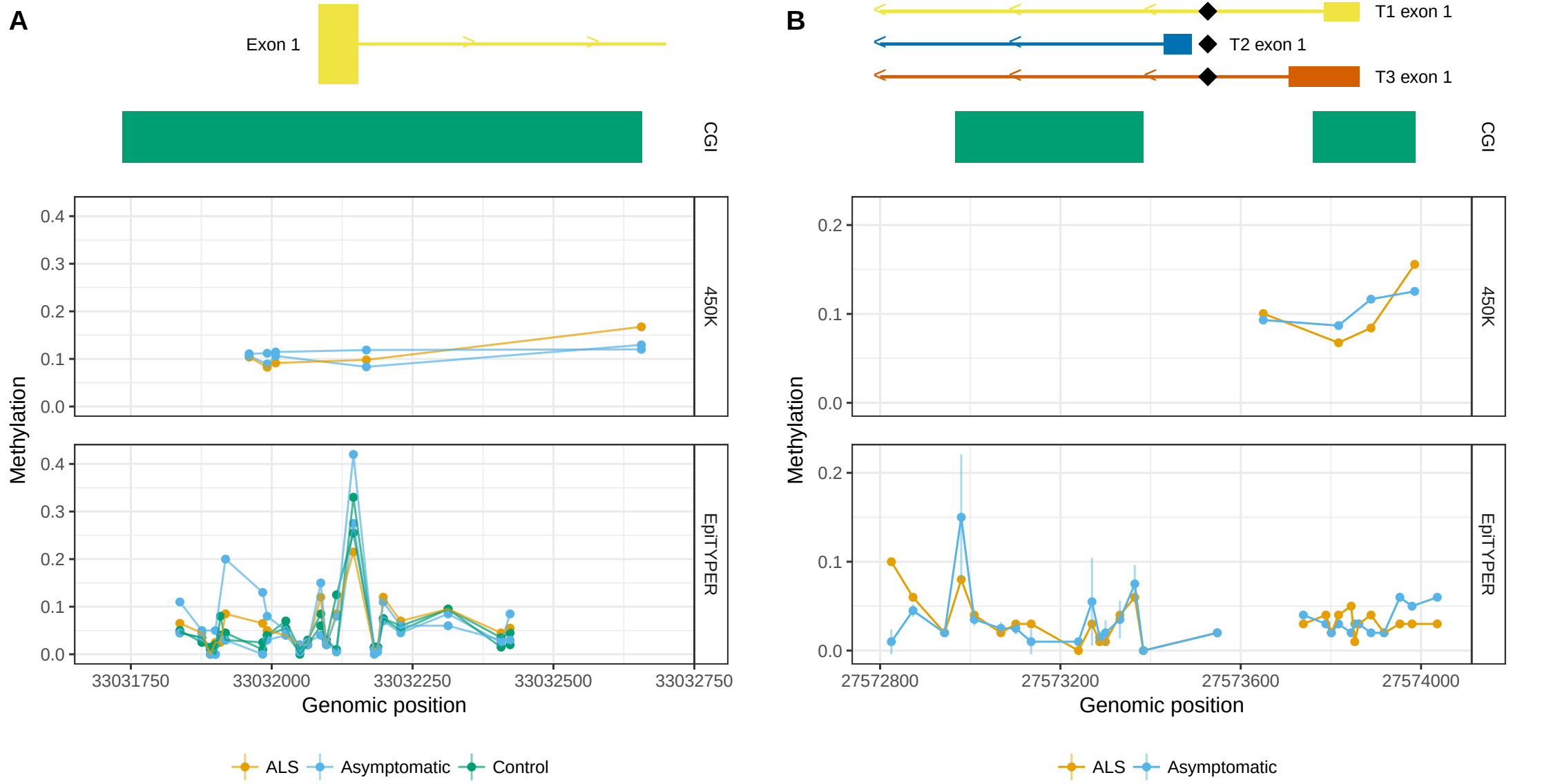
92. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J. 2017;9(2):378-400.

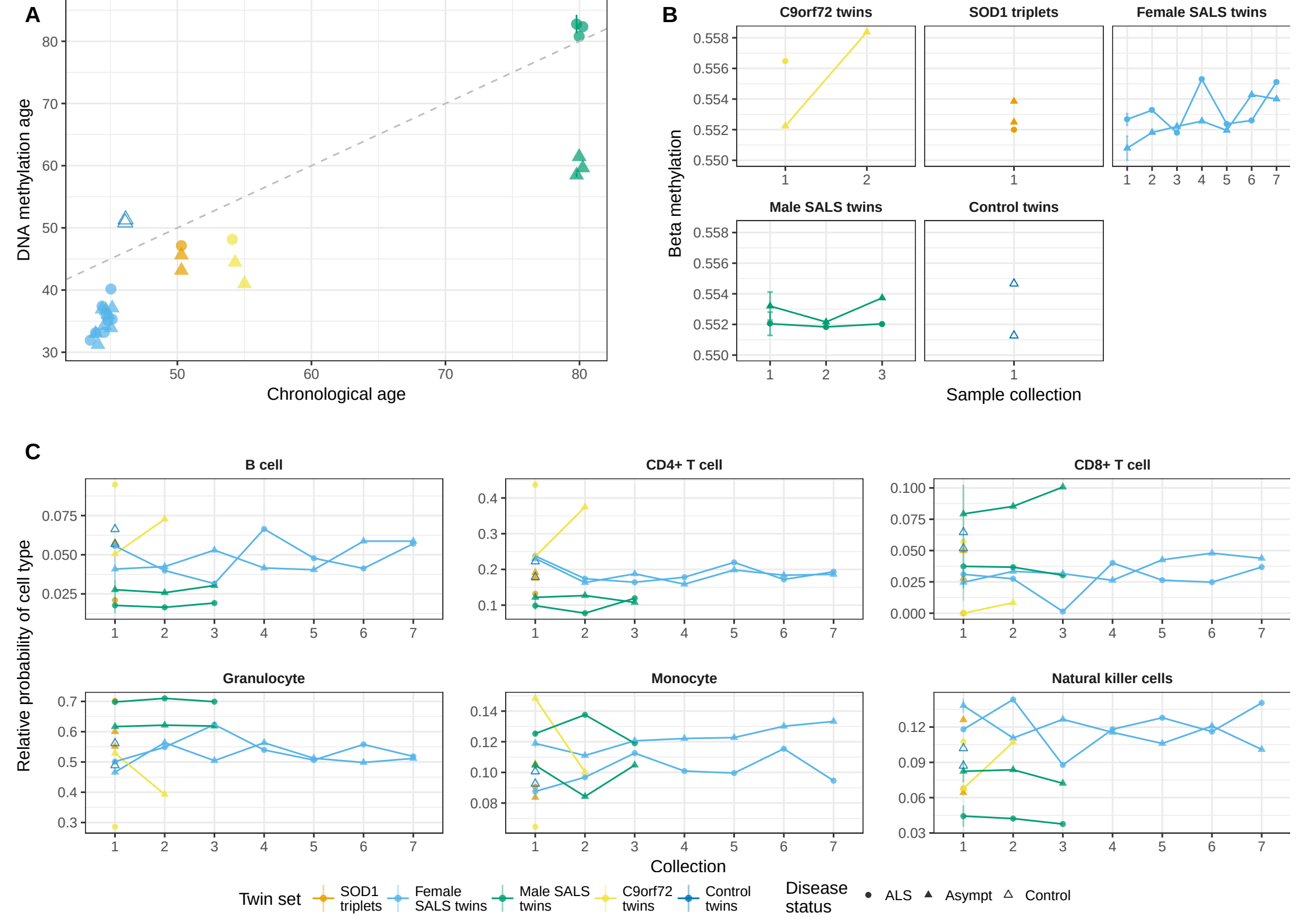
93. Clifford H, Wessely F, Pendurthi S, Emes RD. Comparison of clustering methods for investigation of genome-wide methylation array data. Front Genet. 2011;2:88.

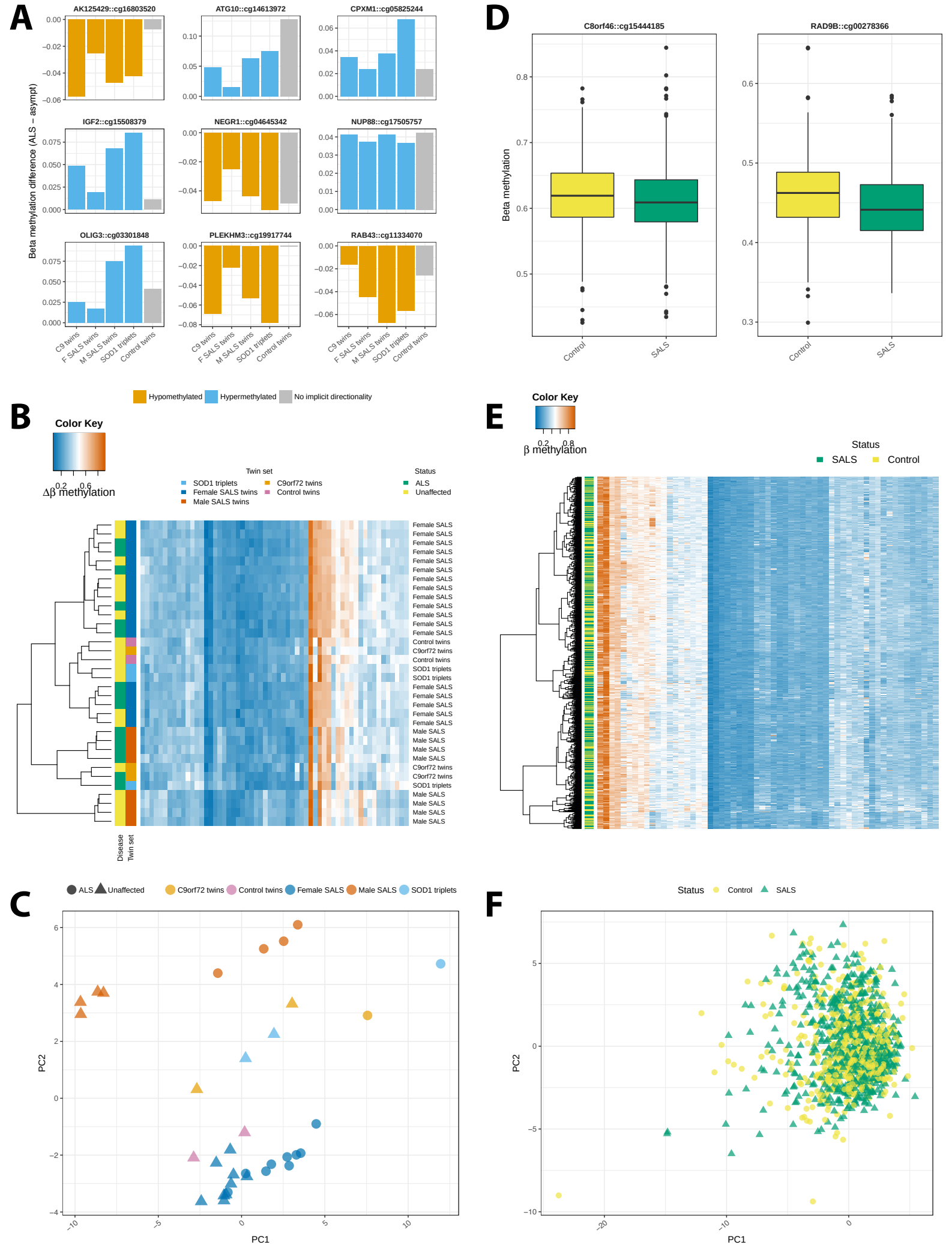
94. Jaskowiak PA, Costa IG, Campello R. Clustering of RNA-Seq samples: Comparison study on cancer data. Methods. 2018;132:42-9.

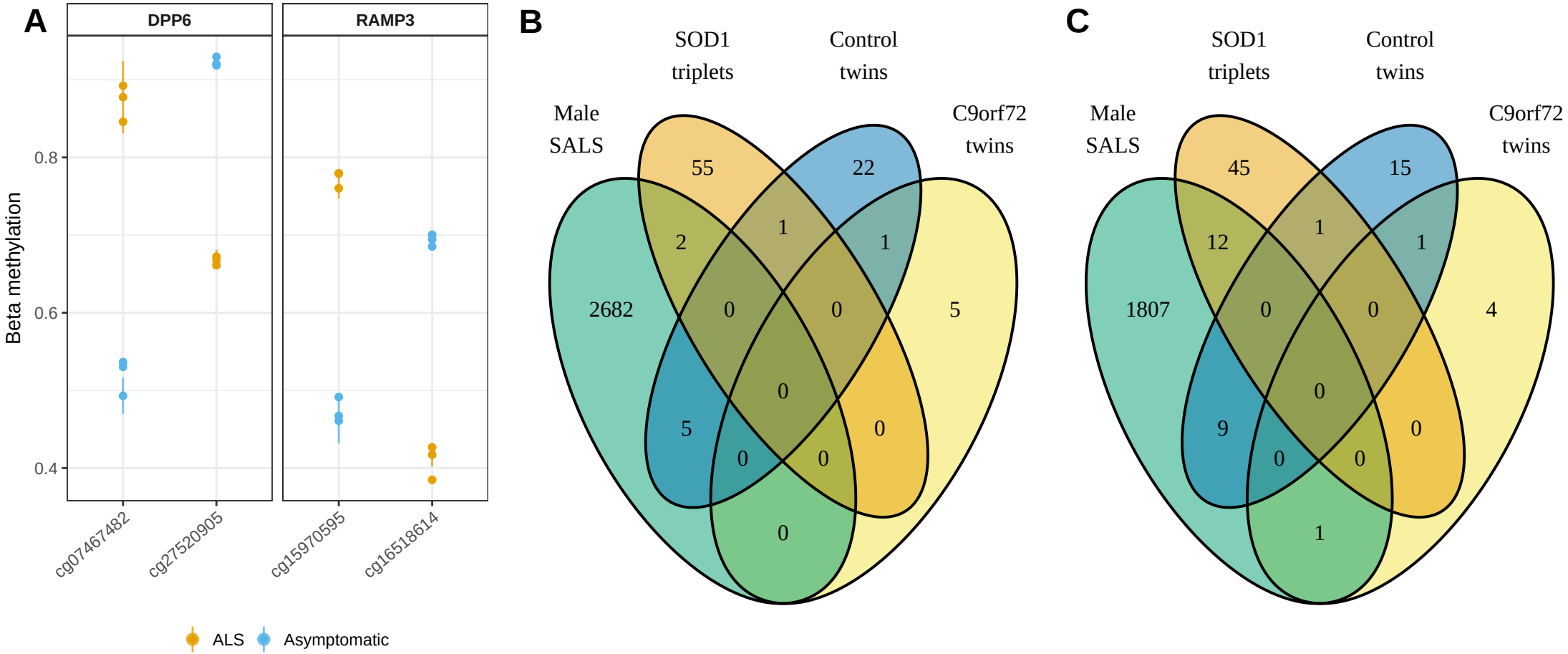
Figures

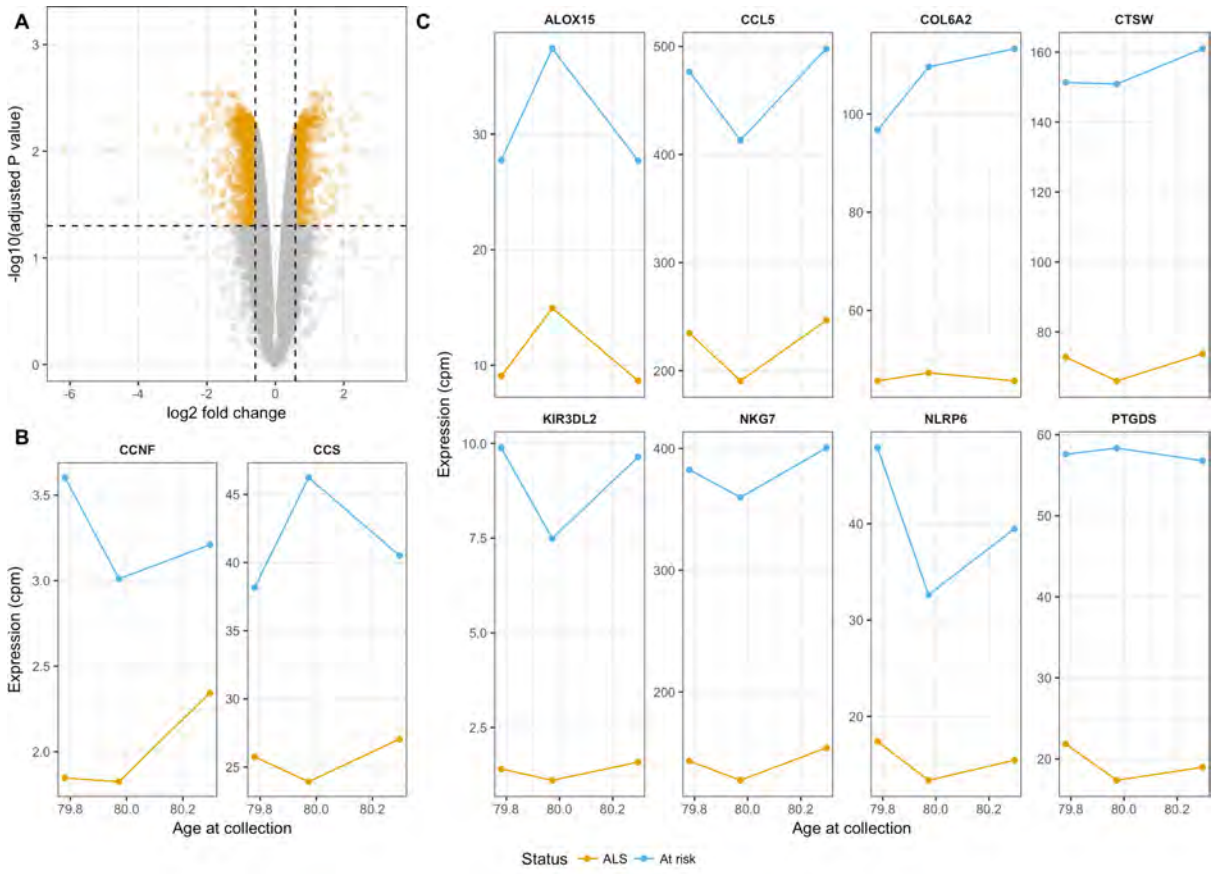

Male SALS twins
Female SALS twins
SOD1 p.I114T
SOD1 p.I114T
SOD1 p.I114T
SOD1 triplets
C9orf72 HRE
C9orf72 HRE
C9orf72 HRE
C9orf72 twins

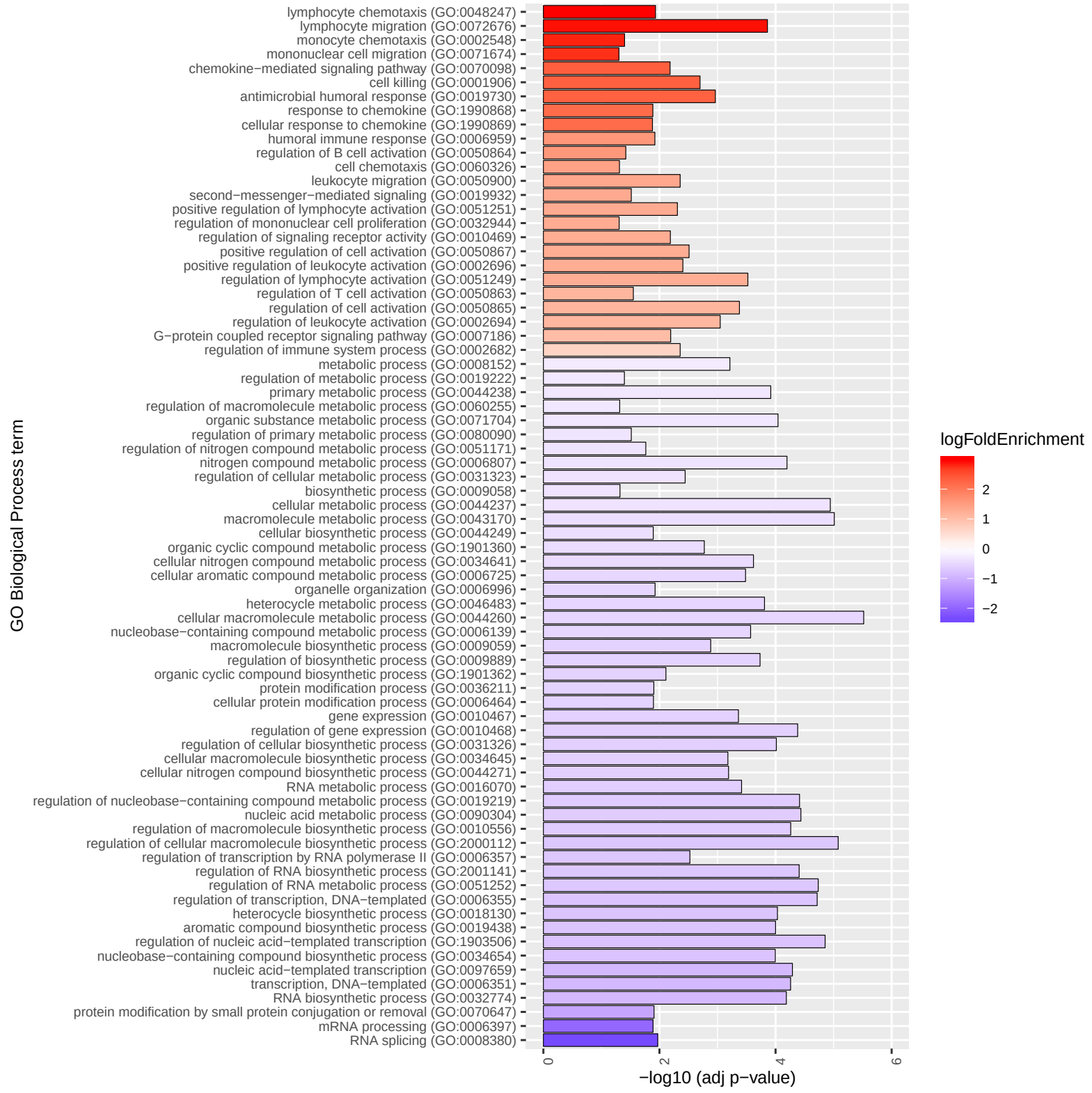

A
Exon 1
CGI
450K
EpiTYPER
Methylation
0.0
0.1
0.2
0.3
0.4
33031750
33032000
33032250
33032500
33032750
Genomic position
ALS
Asymptomatic
Control
B
T1 exon 1
T2 exon 1
T3 exon 1
CGI
450K
EpiTYPER
Methylation
0.0
0.1
0.2
27572800
27573200
27573600
27574000
Genomic position
ALS
Asymptomatic

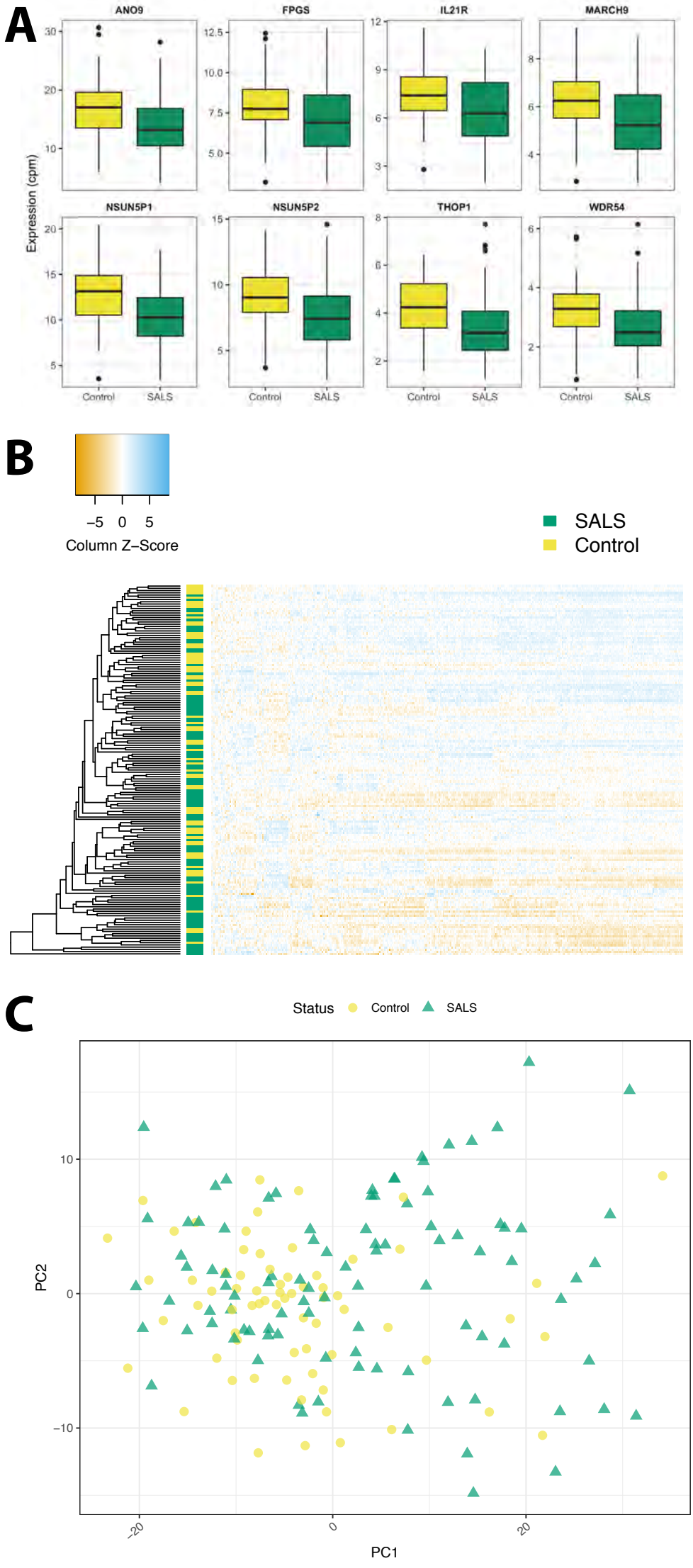

A
DNA methylation age
Chronological age
B
C9orf72 twins
SOD1 triplets
Female SALS twins
Male SALS twins
Control twins
Beta methylation
Sample collection
C
B cell
CD4+ T cell
CD8+ T cell
Granulocyte
Monocyte
Natural killer cells
Relative probability of cell type
Collection
Twin set
SOD1 triplets
Female SALS twins
Male SALS twins
C9orf72 twins
Control twins
Disease status
ALS
Asympt
Control

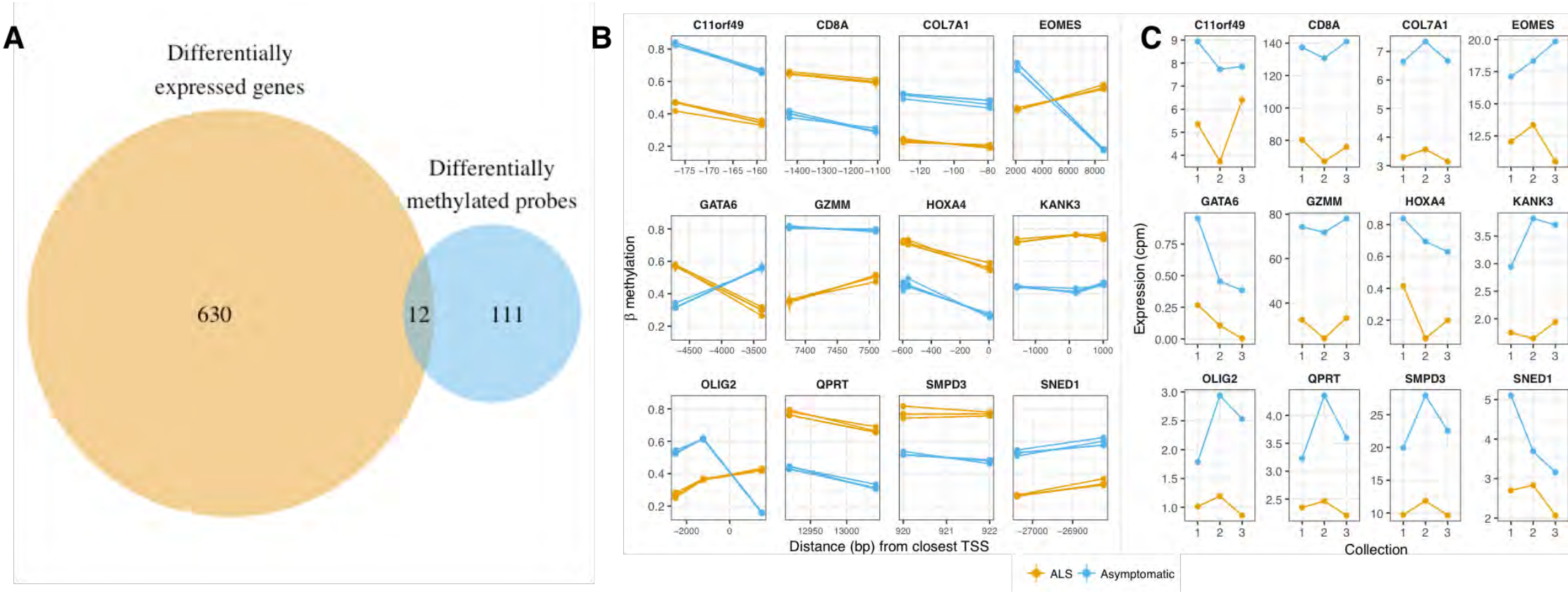

A
AK125429::cg16803520
ATG10::cg14613972
CPXM1::cg05825244
IGF2::cg15508379
NEGR1::cg04645342
NUP88::cg17505757
OLIG3::cg03301848
PLEKHM3::cg19917744
RAB43::cg11334070
Beta methylation difference (ALS − asympt)
C9 twins
F SALS twins
M SALS twins
SOD1 triplets
Control twins
Hypomethylated
Hypermethylated
No implicit directionality
B
Color Key
Δβ methylation
Twin set
SOD1 triplets
Female SALS twins
Male SALS twins
C9orf72 twins
Control twins
Status
ALS
Unaffected
Female SALS
Male SALS
Control twins
C9orf72 twins
SOD1 triplets
Disease
Twin set
C
ALS
Unaffected
C9orf72 twins
Control twins
Female SALS
Male SALS
SOD1 triplets
PC1
PC2
D
C8orf46::cg15444185
RAD9B::cg00278366
Beta methylation
Control
SALS
E
Color Key
β methylation
Status
SALS
Control
F
Status
Control
SALS
PC1
PC2


A
DPP6
RAMP3
Beta methylation
0.8
0.6
0.4
cg07467482
cg27520905
cg15970595
cg16518614
ALS
Asymptomatic
B
SOD1 triplets
Control twins
Male SALS
C9orf72 twins
55
22
2
1
1
2682
0
0
5
0
5
0
0
0
0
C
SOD1 triplets
Control twins
Male SALS
C9orf72 twins
45
15
12
1
1
1807
0
0
4
0
9
0
0
0
1


A
-log10(adjusted P value)
log2 fold change
B
CCNF
CCS
Expression (cpm)
Age at collection
C
ALOX15
CCL5
COL6A2
CTSW
KIR3DL2
NKG7
NLRP6
PTGDS
Expression (cpm)
Age at collection
Status
ALS
At risk


lymphocyte chemotaxis (GO:0048247)
lymphocyte migration (GO:0072676)
monocyte chemotaxis (GO:0002548)
mononuclear cell migration (GO:0071674)
chemokine−mediated signaling pathway (GO:0070098)
cell killing (GO:0001906)
antimicrobial humoral response (GO:0019730)
response to chemokine (GO:1990868)
cellular response to chemokine (GO:1990869)
humoral immune response (GO:0006959)
regulation of B cell activation (GO:0050864)
cell chemotaxis (GO:0060326)
leukocyte migration (GO:0050900)
second−messenger−mediated signaling (GO:0019932)
positive regulation of lymphocyte activation (GO:0051251)
regulation of mononuclear cell proliferation (GO:0032944)
regulation of signaling receptor activity (GO:0010469)
positive regulation of cell activation (GO:0050867)
positive regulation of leukocyte activation (GO:0002696)
regulation of lymphocyte activation (GO:0051249)
regulation of T cell activation (GO:0050863)
regulation of cell activation (GO:0050865)
regulation of leukocyte activation (GO:0002694)
G−protein coupled receptor signaling pathway (GO:0007186)
regulation of immune system process (GO:0002682)
metabolic process (GO:0008152)
regulation of metabolic process (GO:0019222)
primary metabolic process (GO:0044238)
regulation of macromolecule metabolic process (GO:0060255)
organic substance metabolic process (GO:0071704)
regulation of primary metabolic process (GO:0080090)
regulation of nitrogen compound metabolic process (GO:0051171)
nitrogen compound metabolic process (GO:0006807)
regulation of cellular metabolic process (GO:0031323)
biosynthetic process (GO:0009058)
cellular metabolic process (GO:0044237)
macromolecule metabolic process (GO:0043170)
cellular biosynthetic process (GO:0044249)
organic cyclic compound metabolic process (GO:1901360)
cellular nitrogen compound metabolic process (GO:0034641)
cellular aromatic compound metabolic process (GO:0006725)
organelle organization (GO:0006996)
heterocycle metabolic process (GO:0046483)
cellular macromolecule metabolic process (GO:0044260)
nucleobase−containing compound metabolic process (GO:0006139)
macromolecule biosynthetic process (GO:0009059)
regulation of biosynthetic process (GO:0009889)
organic cyclic compound biosynthetic process (GO:1901362)
protein modification process (GO:0036211)
cellular protein modification process (GO:0006464)
gene expression (GO:0010467)
regulation of gene expression (GO:0010468)
regulation of cellular biosynthetic process (GO:0031326)
cellular macromolecule biosynthetic process (GO:0034645)
cellular nitrogen compound biosynthetic process (GO:0044271)
RNA metabolic process (GO:0016070)
regulation of nucleobase−containing compound metabolic process (GO:0019219)
nucleic acid metabolic process (GO:0090304)
regulation of macromolecule biosynthetic process (GO:0010556)
regulation of cellular macromolecule biosynthetic process (GO:2000112)
regulation of transcription by RNA polymerase II (GO:0006357)
regulation of RNA biosynthetic process (GO:2001141)
regulation of RNA metabolic process (GO:0051252)
regulation of transcription, DNA−templated (GO:0006355)
heterocycle biosynthetic process (GO:0018130)
aromatic compound biosynthetic process (GO:0019438)
regulation of nucleic acid−templated transcription (GO:1903506)
nucleobase−containing compound biosynthetic process (GO:0034654)
nucleic acid−templated transcription (GO:0097659)
transcription, DNA−templated (GO:0006351)
RNA biosynthetic process (GO:0032774)
protein modification by small protein conjugation or removal (GO:0070647)
mRNA processing (GO:0006397)
RNA splicing (GO:0008380)
GO Biological Process term
0
2
4
6
−log10 (adj p−value)
logFoldEnrichment
2
1
0
−1
−2


A
ANO9
FPGS
IL21R
MARCH9
NSUN5P1
NSUN5P2
THOP1
WDR54
Expression (cpm)
Control
SALS
B
−5 0 5
Column Z−Score
SALS
Control
C
Status
Control
SALS
PC2
PC1
−20
0
20
10
0
−10


A
Differentially expressed genes
Differentially methylated probes
630
12
111
B
C11orf49
CD8A
COL7A1
EOMES
GATA6
GZMM
HOXA4
KANK3
OLIG2
QPRT
SMPD3
SNED1
β methylation
Distance (bp) from closest TSS
C
Expression (cpm)
Collection
ALS
Asymptomatic

References

S. Abdelkarim, S. Morgan, V. Plagnol, C. H. Lu, G. Adamson, R. Howard, A. Malaspina, R. Orrell, N. Sharma, K. Sidle, J. Clarke, N. C. Fox, M. N. Rossor, J. D. Warren, C. N. Clark, J. D. Rohrer, E. M. Fisher, S. Mead, A. Pittman, and P. Fratta. CHCHD10 Pro34Ser is not a highly penetrant pathogenic variant for amyotrophic lateral sclerosis and frontotemporal dementia. *Brain*, 139(Pt 2):e9, Feb 2016. 106

K. Abe, M. Aoki, S. Tsuji, Y. Itoyama, G. Sobue, M. Togo, C. Hamada, M. Tanaka, M. Akimoto, K. Nakamura, F. Takahashi, K. Kondo, H. Yoshino, K. Abe, M. Aoki, S. Tsuji, Y. Itoyama, G. Sobue, M. Togo, C. Hamada, H. Sasaki, I. Yabe, S. Doi, H. Warita, T. Imai, H. Ito, M. Fukuchi, E. Osumi, M. Wada, I. Nakano, M. Morita, K. Ogata, Y. Maruki, K. Ito, O. Kano, M. Yamazaki, Y. Takahashi, H. Ishiura, M. Ogino, R. Koike, C. Ishida, T. Uchiyama, K. Mizoguchi, T. Obi, H. Watanabe, N. Atsuta, I. Aiba, A. Taniguchi, H. Sawada, T. Hazama, H. Fujimura, H. Kusaka, T. Kunieda, H. Kikuchi, H. Matsuo, H. Ueyama, K. Uekawa, M. Tanaka, M. Akimoto, M. Ueda, A. Murakami, R. Sumii, T. Kudou, K. Nakamura, K. Morimoto, T. Yoneoka, M. Hirai, K. Sasaki, H. Terai, T. Natori, H. Matsui, K. Kotani, K. Yoshida, T. Iwasaki, F. Takahashi, K. Kondo, and H. Yoshino. Safety and efficacy of edaravone in well defined patients with amyotrophic lateral sclerosis: a randomised, double-blind, placebo-controlled trial. *Lancet Neurol*, 16(7):505–512, Jul 2017. 7

G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, 30(1):97–101, Jan 2002. 64, 180

O. Abel, J. F. Powell, P. M. Andersen, and A. Al-Chalabi. ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.*, 33(9):1345–1351, Sep 2012. 14

O. Abel, A. Shatunov, A. R. Jones, P. M. Andersen, J. F. Powell, and A. Al-Chalabi. Development of a Smartphone App for a Genetics Website: The Amyotrophic Lateral Sclerosis Online Genetics Database (ALSoD). *JMIR Mhealth Uhealth*, 1(2):e18, 2013. 12

A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, 21(6):974–984, Jun 2011. 305

I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat. Methods*, 7(4):248–249, Apr 2010. 57

D. Aird, M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, 12(2):R18, 2011. 270, 292

O. Aitio, M. Hellman, A. Kazlauskas, D. F. Vingadassalom, J. M. Leong, K. Saksela, and P. Permi. Recognition of tandem PxxP motifs as a unique Src homology 3-binding mode triggers pathogen-driven actin assembly. *Proc. Natl. Acad. Sci. U.S.A.*, 107(50):21743–21748, Dec 2010. 24

S. S. Ajay, S. C. Parker, H. O. Abaan, K. V. Fajardo, and E. H. Margulies. Accurate and comprehensive sequencing of personal genomes. *Genome Res.*, 21(9):1498–1505, Sep 2011. 294

A. Al-Chalabi and O. Hardiman. The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat Rev Neurol*, 9(11):617–628, Nov 2013. 3, 14, 15, 303

A. Al-Chalabi, F. Fang, M. F. Hanby, P. N. Leigh, C. E. Shaw, W. Ye, and F. Rijsdijk. An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatr.*, 81(12):1324–1326, Dec 2010. 26, 223

A. Al-Chalabi, A. Calvo, A. Chio, S. Colville, C. M. Ellis, O. Hardiman, M. Heverin, R. S. Howard, M. H. B. Huisman, N. Keren, P. N. Leigh, L. Mazzini, G. Mora, R. W. Orrell, J. Rooney, K. M. Scott, W. J. Scotton, M. Seelen, C. E. Shaw, K. S. Sidle, R. Swingler, M. Tsuda, J. H. Veldink, A. E. Visser, L. H. van den Berg, and N. Pearce. Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *Lancet Neurol*, 13(11):1108–1113, Nov 2014. 271, 284

A. Al-Chalabi, L. H. van den Berg, and J. Veldink. Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nat Rev Neurol*, 13(2):96–104, Feb 2017. 10, 13, 25, 26, 27, 37, 278

S. Almeida, E. Gascon, H. Tran, H. J. Chou, T. F. Gendron, S. Degroot, A. R. Tapper, C. Sellier, N. Charlet-Berguerand, A. Karydas, W. W. Seeley, A. L. Boxer, L. Petrucelli, B. L. Miller, and F. B. Gao. Modeling key pathological features of frontotemporal dementia with C9ORF72 repeat expansion in iPSC-derived human neurons. *Acta Neuropathol.*, 126(3):385–399, Sep 2013. 21

D. Altshuler, M. J. Daly, and E. S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, Nov 2008. 29, 30

P. M. Andersen. Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Curr Neurol Neurosci Rep*, 6(1):37–46, Jan 2006. 92, 143

P. M. Andersen and A. Al-Chalabi. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nat Rev Neurol*, 7(11):603–615, Nov 2011. 11, 13, 14, 15, 18, 303

P. M. Andersen, L. Forsgren, M. Binzer, P. Nilsson, V. Ala-Hurula, M. L. Keranen, L. Bergmark, A. Saarinen, T. Haltia, I. Tarvainen, E. Kinnunen, B. Udd, and S. L. Marklund. Autosomal recessive adult-onset amyotrophic lateral sclerosis associated with homozygosity for Asp90Ala CuZn-superoxide dismutase mutation. A clinical and genealogical study of 36 patients. *Brain*, 119 (Pt 4):1153–1172, Aug 1996. 14

T. Arai, M. Hasegawa, H. Akiyama, K. Ikeda, T. Nonaka, H. Mori, D. Mann, K. Tsuchiya, M. Yoshida, Y. Hashizume, and T. Oda. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochem. Biophys. Res. Commun.*, 351(3):602–611, Dec 2006. 168

P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer*, 8(1):1–12, Mar 1954. 284

P. E. Ash, K. F. Bieniek, T. F. Gendron, T. Caulfield, W. L. Lin, M. Dejesus-Hernandez, M. M. van Blitterswijk, K. Jansen-West, J. W. Paul, R. Rademakers, K. B. Boylan, D. W. Dickson, and L. Petrucelli. Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ALS. *Neuron*, 77(4):639–646, Feb 2013. 9, 22

N. authors listed. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, 72(6):971–983, Mar 1993. 283, 305

A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. J. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. Del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll,

J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. Stutz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, G. R. Abecasis, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, R. M. Durbin, M. E. Hurles, C. Tyler-Smith, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W. P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. Stutz, P. Flicek,

F. P. Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley, B. Barnes, R. K. Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E. W. Lameijer, M. A. Batzer, M. K. Konkel, J. A. Walker, L. Ding, I. Hall, K. Ye, P. Lacroute, C. Lee, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine, E. J. Gardner, G. R. Abecasis, J. M. Kidd, R. E. Mills, G. Dayama, S. Emery, G. Jun, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, K. Chen, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana, M. E. Hurles, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. J. Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth, E. P. Garrison, W. F. Leong, A. N. Ward, G. Del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur, E. R. Mardis, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E. T. Dermitzakis, G. R. Abecasis, H. Min Kang, G. A. McVean, M. B. Gerstein, S. Balasubramanian, L. Habegger, M. B. Gerstein, S. Balasubramanian, L. Habegger, H. Yu, P. Flicek, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage, J. B. Jespersen, H. Horn, S. B. Montgomery, M. K. DeGorter, E. Khurana, C. Tyler-Smith, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, A. Auton, A. Marcketta, R. Desalle, A. Narechania, M. A. Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. F. Willems, C. D. Bustamante, F. L. Mendez, G. D. Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis, L. Coin, H. Shao, D. Mittelman, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek, L. Clarke, X. Zheng-Bradley, D. R. Bentley, A. Cox, S. Humphray, S. Kahn,

R. Sudbrak, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig, T. Izatt, A. A. Kurdoglu, S. T. Sherry, C. Xiao, D. Haussler, G. R. Abecasis, G. A. McVean, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. Hien, S. J. Dunstan, N. T. Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015. 208

M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, 12(11):745–755, Nov 2011. 45, 286

S. Banfi, A. Servadio, M. Y. Chung, T. J. Kwiatkowski, A. E. McCall, L. A. Duvick, Y. Shen, E. J. Roth, H. T. Orr, and H. Y. Zoghbi. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.*, 7(4):513–520, Aug 1994. 283, 305

S. Bannwarth, S. Ait-El-Mkadem, A. Chaussenot, E. C. Genin, S. Lacas-Gervais, K. Fragaki, L. Berg-Alonso, Y. Kageyama, V. Serre, D. G. Moore, A. Verschueren, C. Rouzier, I. Le Ber, G. Auge, C. Cochaud, F. Lespinasse, K. N'Guyen,

A. de Septenville, A. Brice, P. Yu-Wai-Man, H. Sesaki, J. Pouget, and V. Paquis-Flucklinger. A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain*, 137(Pt 8):2329–2345, Aug 2014. 12, 25, 35, 106, 143, 154

S. Bao, R. Jiang, W. Kwan, B. Wang, X. Ma, and Y. Q. Song. Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.*, 56(6): 406–414, Jun 2011. 293

S. E. Baranzini, J. Mudge, J. C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, N. A. Miller, L. Zhang, A. D. Farmer, C. J. Bell, R. W. Kim, G. D. May, J. E. Woodward, S. J. Caillier, J. P. McElroy, R. Gomez, M. J. Pando, L. E. Clendenen, E. E. Ganusova, F. D. Schilkey, T. Ramaraj, O. A. Khan, J. J. Huntley, S. Luo, P. Y. Kwok, T. D. Wu, G. P. Schroth, J. R. Oksenberg, S. L. Hauser, and S. F. Kingsmore. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, 464(7293):1351–1356, Apr 2010. 267, 279

A. Barman, A. Assmann, S. Richter, J. Soch, H. Schutze, T. Wustenberg, A. Deibele, M. Klein, A. Richter, G. Behnisch, E. Duzel, M. Zenker, C. I. Seidenbecher, and B. H. Schott. Genetic variation of the RASGRF1 regulatory region affects human hippocampus-dependent memory. *Front Hum Neurosci*, 8:260, 2014. 217

A. Belkadi, A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko, L. Shang, B. Boisson, J. L. Casanova, and L. Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U.S.A.*, 112(17):5473–5478, Apr 2015. 286, 287, 295

L. Bello, K. M. Flanigan, R. B. Weiss, P. Spitali, A. Aartsma-Rus, F. Muntoni, I. Zaharieva, A. Ferlini, E. Mercuri, S. Tuffery-Giraud, M. Claustres, V. Straub, H. Lochmuller, A. Barp, S. Vianello, E. Pegoraro, J. Punetha, H. Gordish-Dressman, M. Giri, C. M. McDonald, E. P. Hoffman, D. M. Dunn, K. J. Swoboda, E. Gappmaier, M. T. Howard, J. B. Sampson, M. B. Bromberg, R. Butterfield, L. Kerr, A. Pestronk, J. M. Florence, A. Connolly, G. Lopate, P. Golumbek, J. Schierbecker, B. Malkus, R. Renna, C. Siener, R. S. Finkel, C. G. Bonnemann, L. Medne, A. M. Glanzman, J. Flickinger, J. R. Mendell, W. M. King, L. Lowes, L. Alfano, K. D. Mathews, C. Stephan, K. Laubenthal, K. Baldwin, B. Wong, P. Morehart, A. Meyer, J. W. Day, C. E. Naughton, M. Margolis, A. Cnaan, R. T. Abresch, E. K. Henricson, L. P. Morgenroth, T. Duong, V. V. Chidambaranathan, W. D. Biggar, L. C. McAdam, J. Mah, M. Tulinius, R. Leshner, C. T. Rocha, M. Thangarajh, A. Kornberg, M. Ryan, Y. Nevo, A. Dubrovsky, P. R. Clemens, H. Abdel-Hamid, A. M.

Connolly, A. Pestronk, J. Teasley, T. E. Bertorini, K. North, R. Webster, H. Kolski, N. Kuntz, S. Driscoll, J. Carlo, K. Gorni, T. Lotze, J. W. Day, P. Karachunski, and J. B. Bodensteiner. Association Study of Exon Variants in the NF-B and TGF Pathways Identifies CD40 as a Modifier of Duchenne Muscular Dystrophy. *Am. J. Hum. Genet.*, 99(5):1163–1171, Nov 2016. 223

V. V. Belzil, P. O. Bauer, M. Prudencio, T. F. Gendron, C. T. Stetler, I. K. Yan, L. Pregent, L. Daughrity, M. C. Baker, R. Rademakers, K. Boylan, T. C. Patel, D. W. Dickson, and L. Petrucelli. Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood. *Acta Neuropathol.*, 126(6):895–905, Dec 2013. 21

V. V. Belzil, P. O. Bauer, T. F. Gendron, M. E. Murray, D. Dickson, and L. Petrucelli. Characterization of DNA hypermethylation in the cerebellum of c9FTD/ALS patients. *Brain Res.*, Feb 2014. 284

V. V. Belzil, R. B. Katzman, and L. Petrucelli. ALS and FTD: an epigenetic perspective. *Acta Neuropathol.*, 132(4):487–502, Oct 2016. 283

G. Bensimon, L. Lacomblez, and V. Meininger. A controlled trial of riluzole in amyotrophic lateral sclerosis. ALS/Riluzole Study Group. *N. Engl. J. Med.*, 330(9): 585–591, Mar 1994. 6

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson,

N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008. 294

B. Benyamin, J. He, Q. Zhao, J. Gratten, F. Garton, P. J. Leo, Z. Liu, M. Mangelsdorf, A. Al-Chalabi, L. Anderson, T. J. Butler, L. Chen, X. D. Chen, K. Cremin, H. W. Deng, M. Devine, J. Edson, J. A. Fifita, S. Furlong, Y. Y. Han, J. Harris, A. K. Henders, R. L. Jeffree, Z. B. Jin, Z. Li, T. Li, M. Li, Y. Lin, X. Liu, M. Marshall, E. P. McCann, B. J. Mowry, S. T. Ngo, R. Pamphlett, S. Ran, D. C. Reutens, D. B. Rowe, P. Sachdev, S. Shah, S. Song, L. J. Tan, L. Tang, L. H. van den Berg, W. van Rheenen, J. H. Veldink, R. H. Wallace, L. Wheeler, K. L. Williams, J. Wu, X. Wu, J. Yang, W. Yue, Z. H. Zhang, D. Zhang, P. G. Noakes, I. P. Blair, R. D. Henderson, P. A. McCombe, P. M. Visscher, H. Xu, P. F. Bartlett, M. A. Brown, N. R. Wray, and D. Fan. Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus with amyotrophic lateral sclerosis. *Nat Commun*, 8(1):611, 09 2017. 12, 27, 154

K. Beanovi, A. Norremolle, S. J. Neal, C. Kay, J. A. Collins, D. Arenillas, T. Lilja, G. Gaudenzi, S. Manoharan, C. N. Doty, J. Beck, N. Lahiri, E. Portales-Casamar, S. C. Warby, C. Connolly, R. A. De Souza, S. J. Tabrizi, O. Hermanson, D. R. Langbehn, M. R. Hayden, W. W. Wasserman, and B. R. Leavitt. A SNP in the HTT

promoter alters NF-B binding and is a bidirectional genetic modifier of Huntington disease. *Nat. Neurosci.*, 18(6):807–816, Jun 2015. 223

I. P. Blair, K. L. Williams, S. T. Warraich, J. C. Durnall, A. D. Thoeng, J. Manavis, P. C. Blumbergs, S. Vucic, M. C. Kiernan, and G. A. Nicholson. FUS mutations in amyotrophic lateral sclerosis: clinical, pathological, neurophysiological and genetic analysis. *J. Neurol. Neurosurg. Psychiatr.*, 81(6):639–645, Jun 2010. 17, 18

H. Blasco, F. Patin, C. R. Andres, P. Corcia, and P. H. Gordon. Amyotrophic Lateral Sclerosis, 2016: existing therapies and the ongoing search for neuroprotection. *Expert Opin Pharmacother*, 17(12):1669–1682, Aug 2016. 5

N. Blom, S. Gammeltoft, and S. Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, 294(5):1351–1362, Dec 1999. 56, 57

D. I. Boomsma. Twin, association and current "omics" studies. *J. Matern. Fetal. Neonatal. Med.*, 26 Suppl 2:9–12, Oct 2013. 34, 222

I. B. Borecki and M. A. Province. Linkage and association: basic concepts. *Adv. Genet.*, 60:51–74, 2008. 29

A. L. Boxer, I. R. Mackenzie, B. F. Boeve, M. Baker, W. W. Seeley, R. Crook, H. Feldman, G. Y. Hsiung, N. Rutherford, V. Laluz, J. Whitwell, D. Foti, E. McDade, J. Molano, A. Karydas, A. Wojtas, J. Goldman, J. Mirsky, P. Sengdy, S. Dearmond, B. L. Miller, and R. Rademakers. Clinical, neuroimaging and neuropathological features of a new chromosome 9p-linked FTD-ALS family. *J. Neurol. Neurosurg. Psychiatr.*, 82(2):196–203, Feb 2011. 19

K. Boylan. Familial Amyotrophic Lateral Sclerosis. *Neurol Clin*, 33(4):807–830, Nov 2015. 13, 14, 15, 16, 18, 20, 22, 23, 24

V. Bozzoni, O. Pansarasa, L. Diamanti, G. Nosari, C. Cereda, and M. Ceroni. Amyotrophic lateral sclerosis and environmental factors. *Funct. Neurol.*, 31(1):7–19, 2016. 6

J. Breckpot, B. Thienpont, M. Bauters, L. C. Tranchevent, M. Gewillig, K. Allegaert, J. R. Vermeesch, Y. Moreau, and K. Devriendt. Congenital heart defects in a novel recurrent 22q11.2 deletion harboring the genes CRKL and MAPK1. *Am. J. Med. Genet. A*, 158A(3):574–580, Mar 2012. 34, 271

W. Brockman, P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ, E. S. Lander, C. Nusbaum, and D. B. Jaffe. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*, 18(5):763–770, May 2008. 270, 291

B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Other Motor Neuron Disord.*, 1(5):293–299, Dec 2000. 42

R. H. Brown and A. Al-Chalabi. Amyotrophic Lateral Sclerosis. *N. Engl. J. Med.*, 377 (2):162–172, 07 2017. 5, 6, 7, 8, 9, 21, 35

C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C. Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, and J. P. Dumanski. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.*, 82(3):763–771, Mar 2008. 34, 271

H. M. Bryson, B. Fulton, and P. Benfield. Riluzole. A review of its pharmacodynamic and pharmacokinetic properties and therapeutic potential in amyotrophic lateral sclerosis. *Drugs*, 52(4):549–563, Oct 1996. 6

J. R. Burrell, G. M. Halliday, J. J. Kril, L. M. Ittner, J. Gotz, M. C. Kiernan, and J. R. Hodges. The frontotemporal dementia-motor neuron disease continuum. *Lancet*, 388 (10047):919–931, Aug 2016. 21

S. Byrne, M. Elamin, P. Bede, and O. Hardiman. Absence of consensus in diagnostic criteria for familial neurodegenerative diseases. *J. Neurol. Neurosurg. Psychiatry*, 83 (4):365–367, Apr 2012a. 37

S. Byrne, M. Elamin, P. Bede, A. Shatunov, C. Walsh, B. Corr, M. Heverin, N. Jordan, K. Kenna, C. Lynch, R. L. McLaughlin, P. M. Iyer, C. O'Brien, J. Phukan, B. Wynne, A. L. Bokde, D. G. Bradley, N. Pender, A. Al-Chalabi, and O. Hardiman. Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet Neurol*, 11(3): 232–240, Mar 2012b. 20

R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, 30(8):1237–1244, Aug 2009. 57

L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, P. Guan, G. E. Korzeniewski, N. C. Lockhart, C. A. Rabiner, A. K. Rao, K. L. Robinson, N. V. Roche, S. J. Sawyer, A. V. Segre, C. E. Shive, A. M. Smith, L. H. Sobin, A. H. Undale, K. M. Valentino, J. Vaught, T. R. Young, H. M. Moore, L. Barker, M. Basile, A. Battle, J. Boyer, D. Bradbury, J. P. Bridge, A. Brown, R. Burges, C. Choi, D. Colantuoni, N. Cox, E. T. Dermitzakis, L. K. Derr, M. J. Dinsmore, K. Erickson, J. Fleming, T. Flutre, B. A. Foster, E. R. Gamazon, G. Getz, B. M. Gillard, R. Guigo, K. W. Hambright, P. Hariharan, R. Hasz, H. K. Im, S. Jewell, E. Karasik, M. Kellis, P. Kheradpour, S. Koester, D. Koller, A. Konkashbaev, T. Lappalainen, R. Little, J. Liu, E. Lo, J. T. Lonsdale, C. Lu, D. G. MacArthur, H. Magazine, J. B. Maller, Y. Marcus, D. C. Mash, M. I. McCarthy, J. McLean, B. Mestichelli, M. Miklos, J. Monlong, M. Mosavel, M. T. Moser, S. Mostafavi, D. L. Nicolae, J. Pritchard, L. Qi, K. Ramsey, M. A. Rivas, B. E. Robles, D. C. Rohrer, M. Salvatore, M. Sammeth, J. Seleski, S. Shad, L. A. Siminoff, M. Stephens, J. Struewing, T. Sullivan, S. Sullivan, J. Syron, D. Tabor, M. Taherian, J. Tejada, G. F. Temple, J. A. Thomas, A. W. Thomson, D. Tidwell, H. M. Traino, Z. Tu, D. R. Valley, S. Volpi, G. D. Walters, L. D. Ward, X. Wen, W. Winckler, S. Wu, and J. Zhu. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank*, 13(5):311–319, Oct 2015. 55, 57

F. Casals, Y. Idaghdour, J. Hussin, and P. Awadalla. Next-generation sequencing approaches for genetic mapping of complex diseases. *J. Neuroimmunol.*, 248(1-2): 10–22, Jul 2012. 45

C. A. Castellani, M. G. Melka, J. L. Gui, A. J. Gallo, R. L. O'Reilly, and S. M. Singh. Post-zygotic genomic changes in glutamate and dopamine pathway genes may explain discordance of monozygotic twins for schizophrenia. *Clin Transl Med*, 6(1):43, Nov 2017. 222

A. Chari, E. Paknia, and U. Fischer. The role of RNP biogenesis in spinal muscular atrophy. *Curr. Opin. Cell Biol.*, 21(3):387–393, Jun 2009. 19

A. Chatr-Aryamontri, B. J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, 43(Database issue):D470–478, Jan 2015. 57

A. Chaussenot, I. Le Ber, S. Ait-El-Mkadem, A. Camuzat, A. de Septenville, S. Bannwarth, E. C. Genin, V. Serre, G. Auge, A. Brice, J. Pouget, and V. Paquis-Flucklinger. Screening of CHCHD10 in a French cohort confirms the involvement of this gene in frontotemporal dementia with amyotrophic lateral sclerosis patients. *Neurobiol. Aging*, 35(12):1–2884, Dec 2014. 106, 107

Y. Z. Chen, C. L. Bennett, H. M. Huynh, I. P. Blair, I. Puls, J. Irobi, I. Dierick, A. Abel, M. L. Kennerson, B. A. Rabin, G. A. Nicholson, M. Auer-Grumbach, K. Wagner, P. De Jonghe, J. W. Griffin, K. H. Fischbeck, V. Timmerman, D. R. Cornblath, and P. F. Chance. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am. J. Hum. Genet.*, 74(6):1128–1135, Jun 2004. 9, 12, 25

A. S. Chen-Plotkin, V. M. Lee, and J. Q. Trojanowski. TAR DNA-binding protein 43 in neurodegenerative disease. *Nat Rev Neurol*, 6(4):211–220, Apr 2010. 15, 16, 17

A. Chesi, B. T. Staahl, A. Jovicic, J. Couthouis, M. Fasolino, A. R. Raphael, T. Yamazaki, L. Elias, M. Polak, C. Kelly, K. L. Williams, J. A. Fifita, N. J. Maragakis, G. A. Nicholson, O. D. King, R. Reed, G. R. Crabtree, I. P. Blair, J. D. Glass, and A. D. Gitler. Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nat. Neurosci.*, 16(7):851–855, Jul 2013. 12, 223

B. A. Chestnut, Q. Chang, A. Price, C. Lesuisse, M. Wong, and L. J. Martin. Epigenetic regulation of motor neuron cell death through DNA methylation. *J. Neurosci.*, 31 (46):16619–16636, Nov 2011. 272, 284

R. Chia, A. Chio, and B. J. Traynor. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *Lancet Neurol*, 17(1):94–102, Jan 2018. 13

A. Chio, B. J. Traynor, F. Lombardo, M. Fimognari, A. Calvo, P. Ghiglione, R. Mutani, and G. Restagno. Prevalence of SOD1 mutations in the Italian ALS population. *Neurology*, 70(7):533–537, Feb 2008. 13

A. Chio, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, A. Al-Chalabi, P. Couratier, V. Drory, J. Esteban, E. Herrero-Hernandez, N. Leigh, M. Leone, A. Calvo, A. Millul, T. Salas, V. Skvortsova, and Z. Stevic. Prognostic factors in ALS: A critical review. *Amyotroph Lateral Scler*, 10(5-6):310–323, 2009a. 272

A. Chio, G. Restagno, M. Brunetti, I. Ossola, A. Calvo, G. Mora, M. Sabatelli, M. R. Monsurro, S. Battistini, J. Mandrioli, F. Salvi, R. Spataro, J. Schymick, B. J.

Traynor, V. La Bella, F. Giannini, C. Ricci, C. Moglia, F. Lombardo, L. Sbaiz, S. Cammarosano, G. Tedeschi, P. Sola, I. Bartolomei, K. Marinou, L. Papetti, A. Conte, M. Luigetti, P. Paladino, C. Caponnetto, and G. Siciliano. Two Italian kindreds with familial amyotrophic lateral sclerosis due to FUS mutation. *Neurobiol. Aging*, 30(8):1272–1275, Aug 2009b. 18

A. Chio, G. Borghero, M. Pugliatti, A. Ticca, A. Calvo, C. Moglia, R. Mutani, M. Brunetti, I. Ossola, M. G. Marrosu, M. R. Murru, G. Floris, A. Cannas, L. D. Parish, P. Cossu, Y. Abramzon, J. O. Johnson, M. A. Nalls, S. Arepalli, S. Chong, D. G. Hernandez, B. J. Traynor, G. Restagno, S. Battistini, F. Giannini, C. Ricci, A. Canosa, S. Gallo, M. R. Monsurro, G. Tedeschi, J. Mandrioli, P. Sola, F. Salvi, I. Bartolomei, G. Mora, K. Marinou, L. Papetti, A. Conte, M. Sabatelli, M. Luigetti, R. Spataro, V. La Bella, P. Paladino, C. Caponnetto, and P. Volanti. Large proportion of amyotrophic lateral sclerosis cases in Sardinia due to a single founder mutation of the TARDBP gene. *Arch. Neurol.*, 68(5):594–598, May 2011a. 16

A. Chio, A. Calvo, C. Moglia, I. Ossola, M. Brunetti, L. Sbaiz, S. L. Lai, Y. Abramzon, B. J. Traynor, and G. Restagno. A de novo missense mutation of the FUS gene in a "true" sporadic ALS case. *Neurobiol. Aging*, 32(3):23–26, Mar 2011b. 18

A. Chio, A. Calvo, L. Mazzini, R. Cantello, G. Mora, C. Moglia, L. Corrado, S. D'Alfonso, E. Majounie, A. Renton, F. Pisano, I. Ossola, M. Brunetti, B. J. Traynor, and G. Restagno. Extensive genetics of ALS: a population-based study in Italy. *Neurology*, 79(19):1983–1989, Nov 2012. 15

A. Chio, L. Mazzini, S. D'Alfonso, L. Corrado, A. Canosa, C. Moglia, U. Manera, E. Bersano, M. Brunetti, M. Barberis, J. H. Veldink, L. H. van den Berg, N. Pearce, W. Sproviero, R. McLaughlin, A. Vajda, O. Hardiman, J. Rooney, G. Mora, A. Calvo, and A. Al-Chalabi. The multistep hypothesis of ALS revisited: The role of genetic mutations. *Neurology*, 91(7):e635–e642, Aug 2018. 271, 284, 285

Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, 7(10):e46688, 2012. 57, 211

C. Y. Chow, J. E. Landers, S. K. Bergren, P. C. Sapp, A. E. Grant, J. M. Jones, L. Everett, G. M. Lenk, D. M. McKenna-Yasek, L. S. Weisman, D. Figlewicz, R. H. Brown, and M. H. Meisler. Deleterious variants of FIG4, a phosphoinositide phosphatase, in patients with ALS. *Am. J. Hum. Genet.*, 84(1):85–88, Jan 2009. 10, 12, 25, 148

L. Christiansen, A. Lenart, Q. Tan, J. W. Vaupel, A. Aviv, M. McGue, and K. Christensen. DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell*, 15(1):149–154, Feb 2016. 272

M. Y. Chung, Y. C. Lu, N. C. Cheng, and B. W. Soong. A novel autosomal dominant spinocerebellar ataxia (SCA22) linked to chromosome 1p21-q23. *Brain*, 126(Pt 6): 1293–1299, Jun 2003. 32

P. Cingolani, V. Patel, M. Coon, T. Nguyen, S. Land, D. Ruden, and X. Lu. Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Frontiers in Genetics*, 3, 2012. 64

E. T. Cirulli, B. N. Lasseigne, S. Petrovski, P. C. Sapp, P. A. Dion, C. S. Leblond, J. Couthouis, Y. F. Lu, Q. Wang, B. J. Krueger, Z. Ren, J. Keebler, Y. Han, S. E. Levy, B. E. Boone, J. R. Wimbish, L. L. Waite, A. L. Jones, J. P. Carulli, A. G. Day-Williams, J. F. Staropoli, W. W. Xin, A. Chesi, A. R. Raphael, D. McKenna-Yasek, J. Cady, J. M. Vianney de Jong, K. P. Kenna, B. N. Smith, S. Topp, J. Miller, A. Gkazi, A. Al-Chalabi, L. H. van den Berg, J. Veldink, V. Silani, N. Ticozzi, C. E. Shaw, R. H. Baloh, S. Appel, E. Simpson, C. Lagier-Tourenne, S. M. Pulst, S. Gibson, J. Q. Trojanowski, L. Elman, L. McCluskey, M. Grossman, N. A. Shneider, W. K. Chung, J. M. Ravits, J. D. Glass, K. B. Sims, V. M. Van Deerlin, T. Maniatis, S. D. Hayes, A. Ordureau, S. Swarup, J. Landers, F. Baas, A. S. Allen, R. S. Bedlack, J. W. Harper, A. D. Gitler, G. A. Rouleau, R. Brown, M. B. Harms, G. M. Cooper, T. Harris, R. M. Myers, D. B. Goldstein, P. C. Sapp, C. S. Leblond, D. McKenna-Yasek, K. P. Kenna, B. N. Smith, S. Topp, J. Miller, A. Gkazi, A. Al-Chalabi, L. H. van den Berg, J. Veldink, V. Silani, N. Ticozzi, J. Landers, F. Baas, C. E. Shaw, J. D. Glass, G. A. Rouleau, R. Brown, O. Hardiman, R. L. McLaughlin, L. Mazzini, I. P. Blair, K. L. Williams, G. A. Nicholson, S. Al-Sarraj, A. King, E. L. Scotter, S. Topp, C. Troakes, C. Vance, S. D'Alfonso, S. Duga, L. Corrado, A. L. ten Asbroek, D. Calini, C. Colombrita, A. Ratti, C. Tiloca, Z. Wu, S. Asress, M. Polak, F. Diekstra, W. van Rheenen, E. W. Danielson, C. Fallini, P. Keagle, E. A. Lewis, J. Kost, G. Soraru, C. Bertolin, G. Querin, B. Castellotti, C. Gellera, V. Pensato, F. Taroni, C. Cereda, S. Gagliardi, M. Ceroni, G. Lauria, J. de Belleroche, G. P. Comi, S. Corti, R. Del Bo, M. R. Turner, K. Talbot, H. Pall, K. E. Morrison, P. J. Shaw, J. Esteban-Perez, A. Garcia-Redondo, and J. L. Munoz-Blanco. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*, 347(6229):1436–1441, Mar 2015. 10, 12, 25, 28, 35, 282

S. Ciura, S. Lattante, I. Le Ber, M. Latouche, H. Tostivint, A. Brice, and E. Kabashi.

Loss of function of C9orf72 causes motor deficits in a zebrafish model of amyotrophic lateral sclerosis. *Ann. Neurol.*, 74(2):180–187, Aug 2013. 21

D. W. Cleveland, N. Laing, P. V. Hurse, and R. H. Brown. Toxic mutants in Charcot's sclerosis. *Nature*, 378(6555):342–343, Nov 1995. 14

A. Conte, S. Lattante, M. Zollino, G. Marangi, M. Luigetti, A. Del Grande, S. Servidei, F. Trombetta, and M. Sabatelli. P525L FUS mutation is consistently associated with a severe form of juvenile amyotrophic lateral sclerosis. *Neuromuscul. Disord.*, 22(1): 73–75, Jan 2012. 18

P. Corcia, P. Valdmanis, S. Millecamps, C. Lionnet, H. Blasco, K. Mouzat, H. Daoud, V. Belzil, R. Morales, N. Pageot, V. Danel-Brunaud, N. Vandenberghe, P. F. Pradat, P. Couratier, F. Salachas, S. Lumbroso, G. A. Rouleau, V. Meininger, and W. Camu. Phenotype and genotype analysis in amyotrophic lateral sclerosis with TARDBP gene mutations. *Neurology*, 78(19):1519–1526, May 2012. 16

J. Couthouis, M. P. Hart, J. Shorter, M. DeJesus-Hernandez, R. Erion, R. Oristano, A. X. Liu, D. Ramos, N. Jethava, D. Hosangadi, J. Epstein, A. Chiang, Z. Diaz, T. Nakaya, F. Ibrahim, H. J. Kim, J. A. Solski, K. L. Williams, J. Mojsilovic-Petrovic, C. Ingre, K. Boylan, N. R. Graff-Radford, D. W. Dickson, D. Clay-Falcone, L. Elman, L. McCluskey, R. Greene, R. G. Kalb, V. M. Lee, J. Q. Trojanowski, A. Ludolph, W. Robberecht, P. M. Andersen, G. A. Nicholson, I. P. Blair, O. D. King, N. M. Bonini, V. Van Deerlin, R. Rademakers, Z. Mourelatos, and A. D. Gitler. A yeast functional screen predicts new candidate ALS disease genes. *Proc. Natl. Acad. Sci. U.S.A.*, 108(52):20881–20890, Dec 2011. 9, 12

J. Couthouis, M. P. Hart, R. Erion, O. D. King, Z. Diaz, T. Nakaya, F. Ibrahim, H. J. Kim, J. Mojsilovic-Petrovic, S. Panossian, C. E. Kim, E. C. Frackelton, J. A. Solski, K. L. Williams, D. Clay-Falcone, L. Elman, L. McCluskey, R. Greene, H. Hakonarson, R. G. Kalb, V. M. Lee, J. Q. Trojanowski, G. A. Nicholson, I. P. Blair, N. M. Bonini, V. M. Van Deerlin, Z. Mourelatos, J. Shorter, and A. D. Gitler. Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 21(13):2899–2911, Jul 2012. 9, 148, 278

M. P. Cox, D. A. Peterson, and P. J. Biggs. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11:485, Sep 2010. 288

M. E. Cudkowicz, D. McKenna-Yasek, P. E. Sapp, W. Chin, B. Geller, D. L. Hayden, D. A. Schoenfeld, B. A. Hosler, H. R. Horvitz, and R. H. Brown. Epidemiology of

mutations in superoxide dismutase in amyotrophic lateral sclerosis. *Ann. Neurol.*, 41 (2):210–221, Feb 1997. 14

F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669, Jan 2015. 294

R. Daber, S. Sukhadia, and J. J. Morrissette. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer Genet*, 206(12):441–448, Dec 2013. 289, 290, 295

S. Dahoun, S. Gagos, M. Gagnebin, C. Gehrig, C. Burgi, F. Simon, C. Vieux, P. Extermann, R. Lyle, M. A. Morris, S. E. Antonarakis, F. Bena, and J. L. Blouin. Monozygotic twins discordant for trisomy 21 and maternal 21q inheritance: a complex series of events. *Am. J. Med. Genet. A*, 146A(16):2086–2093, Aug 2008. 34, 271

M. Dai, R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan, D. M. Markovitz, G. Omenn, and F. Meng. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, 11 Suppl 4:S7, Dec 2010. 288

J. G. Daigle, N. A. Lanson, R. B. Smith, I. Casci, A. Maltare, J. Monaghan, C. D. Nichols, D. Kryndushkin, F. Shewmaker, and U. B. Pandey. RNA-binding ability of FUS regulates neurodegeneration, cytoplasmic mislocalization and incorporation into stress granules associated with FUS carrying ALS-linked mutations. *Hum. Mol. Genet.*, 22(6):1193–1205, Mar 2013. 18, 19

G. M. Dal, B. Erguner, M. S. Sagiroglu, B. Yuksel, O. E. Onat, C. Alkan, and T. Ozcelik. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J. Med. Genet.*, 51(7):455–459, Jul 2014. 222, 267

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, R. Durbin, D. Altshuler, G. Abecasis, D. Bentley, A. Chakravarti, A. Clark, F. De La Vega,

P. Donnelly, M. Dunn, P. Flicek, S. Gabriel, E. Green, R. Gibbs, B. Knoppers, E. Lander, H. Lehrach, E. Mardis, G. Marth, G. McVean, D. Nickerson, J. Schmidt, S. Sherry, J. Wang, R. Wilson, B. Knoppers, A. Chakravarti, G. Abecasis, K. Barnes, G. Bedoya, L. Brooks, S. Bull, E. Burchard, C. Bustamante, N. Clemm, D. Conway, T. Corrah, F. De La Vega, S. Dunstan, J. Dutil, C. Gallo, A. Garcia-Montero, N. Gharani, R. Gibbs, C. Gignoux, S. Gravel, B. Henn, R. Heyderman, S. Humphray, M. Jallow, L. Jorde, J. Kaye, A. Keinan, A. Kent, T. Kumarasamy, D. Kwiatkowski, C. Lee, R. Li, T. Li, J. C. Martinez-Cruzado, R. Mathias, A. McGuire, G. McVean, J. McEwen, A. Moreno, J. Mullikin, T. Oleksyk, P. Ossorio, M. Parker, V. Parra, G. Poletti, D. Presgraves, D. Reich, C. Rotimi, A. Ruiz-Linares, D. Saleheen, F. S. Joof, Y. Su, R. Sudbrak, H. Taylor, B. Timmermann, S. Tishkoff, L. Toji, C. Tyler-Smith, M. Via, W. Wang, J. Wilson, G. Yang, L. Yang, E. Mardis, S. Gabriel, G. Abecasis, L. Ambrogio, S. Balasubramaniam, Z. Belaia, D. Bentley, T. Blackwell, T. Borodina, L. Brooks, K. Cibulskis, L. Clarke, N. Clemm, A. Coffrey, G. Costa, A. Dahl, P. Danecek, F. De La Vega, P. Danecek, R. Durbin, A. Felsenfeld, T. Fennell, P. Flicek, L. Fulton, R. Gibbs, S. Humphray, A. Indap, Z. Iqbal, D. Jaffe, S. Katzman, T. Keane, R. Kuhn, R. Li, S. McCarthy, K. McKernan, G. McVean, M. Metzker, F. Mertes, D. Muzny, E. Nickerson, C. Nusbaum, A. Palotie, J. Reid, E. Shefler, S. Sherry, M. Shumway, R. Smith, A. Soldatov, C. Sougnez, J. Stalker, R. Sudbrak, H. Swerdlow, B. Timmermann, J. Wang, G. Weinstock, R. Wilson, C. Xiao, H. Yang, H. Z. Bradley, G. McVean, G. Abecasis, A. Abyzov, R. Agarwala, K. Albers, C. Alkan, D. Altshuler, V. Amstislavskiy, P. Anderson, G. del Angel, L. Arbiza, A. Auton, P. Awadalla, Q. Ayub, V. Bafna, M. Bainbridge, S. Balasubramaniam, A. Ball, E. Ball, S. Balasubramanian, E. Banks, J. Barrett, A. Bashir, M. Batzer, M. Bauer, D. Bentley, E. Birney, B. Blackburne, T. Blackwell, R. Blekhman, T. Bloom, A. Boyko, P. Bonnen, A. Brisbin, L. Brooks, B. Browning, S. Browning, C. Bustamante, J. Byrnes, A. Cappellieri, M. Caccamo, N. Cardin, A. Chakravarti, D. Challis, K. Cheetham, K. Chen, Y. Chen, W. Chen, A. Chinwalla, A. Christoforides, K. Cibulskis, A. Clark, L. Clarke, N. Clemm, C. Coarfa, A. Coffrey, D. Conrad, T. Cox, D. Craig, M. Curran, M. Daly, P. Danecek, P. de Bakker, J. Degenhardt, F. De La Vega, M. DePristo, M. Dermitzakis, J. Ding, L. Ding, P. Donnelly, D. Dooling, J. Du, J. Durbin, R. Durbin, M. Eberle, E. Eichler, V. Fallon, T. Fennell, P. Flicek, C. Freeman, M. Fromer, Y. Fu, Y. Fu, S. Gabriel, D. Gaffney, K. Garimella, E. Garrison, M. Gerstein, R. Gibbs, S. Gottipati, S. Gravel, D. Greer, S. Grossman, F. Grubert, R. Gutenkunst, G. Ha, L. Habegger, M. Haimel, I. Hajirasouliha, B. Handsaker, M. Hanna, N. Hansen, R. Haraksingh, C. Hartl, D. Haussler,

B. Henn, R. Hernandez, J. Herrero, R. Herwig, A. Hinrichs, N. Homer, F. Hormozdiari, B. Howie, M. Hu, N. Huang, W. Huang, S. Humphray, M. Hurles, F. Hyland, A. Indap, Z. Iqbal, T. Izatt, D. Jaffe, H. Jin, S. Jones, L. Jorde, L. Jostins, M. Kaganovich, S. Kahn, H. M. Kang, J. Kang, S. Katzman, T. Keane, J. Keebler, J. Kelley, J. Kent, A. Keinan, A. Kern, A. Kernytsky, P. Kersey, E. Khurana, J. Kidd, A. Ko, D. Koboldt, M. Konkel, J. Korbel, J. Korn, R. Kuhn, D. Kural, A. Kurdoglu, P. Lacroute, H. Lam, E. Lander, T. Lappalainen, Q. Le, C. Lee, J. Lee, J. Lee, W. P. Lee, J. Leng, W. F. Leong, B. Li, G. Li, H. Li, L. Li, R. Li, Y. Li, Y. Li, L. Liang, K. Lohmueller, J. Long, Q. Long, G. Lunter, X. Ma, D. MacArthur, J. Maguire, P. Majumder, V. Makarov, J. Marchini, E. Mardis, J. Marioni, G. Marth, S. McCarroll, S. McCarthy, A. McKenna, M. McLellan, C. Melton, B. Merriman, R. Mills, S. Montgomery, L. Moutsianas, X. Mu, J. Mullikin, L. Murray, S. Musharoff, D. Muzny, S. Myers, S. Nelson, J. Nemesh, E. Nickerson, R. Nielsen, Z. Ning, D. Parkhomchuk, E. Parkhomenko, J. Paschall, J. Pearson, H. Peckham, R. Poplin, D. Presgraves, J. Pickrell, A. Price, J. Pritchard, M. Przeworski, S. Purcell, A. Quinlan, J. Reid, M. Rivas, J. Rosenfeld, A. Ruiz-Linares, P. Sabeti, A. Sabo, S. Sajjadian, O. Sakarya, A. Scally, S. Schaffner, P. Scheet, J. Sebat, K. Shakir, T. Sharpe, R. Shaw, E. Shefler, S. Sherry, X. M. Shi, I. Shlyakhter, M. Shumway, S. Sinari, A. Sivachenko, R. Smith, M. Snyder, M. Snyder, C. Sougnez, K. Squire, J. Stalker, M. Stephens, C. Stewart, M. Stromberg, Y. Su, R. Sudbrak, P. Sudmant, Y. Sun, P. Taylor, D. Thomas, S. Tishkoff, E. Tsung, C. Tyler-Smith, A. Urban, C. Vincenza, J. Wallis, J. Walker, K. Walter, Z. Wan, J. Wang, W. Wang, Y. Wang, Y. Wang, A. Ward, T. Webster, G. Weinstock, M. Wendl, D. Wheeler, Y. Wong, J. Wu, C. Xiao, J. Xing, Y. Xue, X. Yan, K. Ye, K. Ye, S. Yoon, F. Yu, J. Yu, D. Zerbino, X. Zhan, Y. Zhan, K. Zhang, Q. Zhang, Y. Zhang, H. Zhao, C. K. Zhang, Z. Zhang, H. Z. Bradley, W. Zheng, M. Zody, S. Zoellner, P. Flicek, S. Sherry, G. Abecasis, C. Alkan, S. Balasubramaniam, E. Birney, T. Bloom, C. Brockington, K. Cibulskis, L. Clarke, G. Cochrane, T. Cox, D. Craig, F. De La Vega, D. Dooling, T. Fennell, R. Gibbs, D. Haussler, A. Hinrichs, S. Humphray, Z. Iqbal, R. Jamieson, S. Kahn, S. Katzman, T. Keane, J. Kent, A. Kern, A. Ko, R. Kuhn, G. Li, R. Li, J. Manning, S. McCarthy, G. McVean, M. Metzker, D. Muzny, W. Nietfeld, J. Paschal, H. Peckham, J. Reid, S. Sajjadian, R. Sanders, S. Sherry, M. Shumway, T. Skelly, R. Smith, J. Stalker, R. Sudbrak, J. Wang, G. Weinstock, J. Wilkinson, C. Xiao, H. Z. Bradley, V. Amstislavskiy, D. Deiros, C. Freeman, G. McVean, M. Przeworski, M. Przeworski, G. Abecasis, L. Arbiza, A. Auton, Q. Ayub, S. Balasubramanian, J. Barrett, R. Blekham, A. Boyko, C. Bustamante, J. Byrnes, A. Chakravarti, A. Clark, D. Conrad, D. Craig, M. Dermitzakis, R. Durbin, Y. Fu, S. Gabriel,

E. Garrison, M. Gerstein, S. Gottipati, S. Gravel, S. Grossman, R. Gutenkunst, L. Habegger, B. Handsaker, B. Henn, R. Hernandez, B. Howie, M. Hu, N. Huang, M. Hurles, Z. Iqbal, A. Indap, H. Jin, L. Jostins, J. Kang, S. Katzman, A. Keinan, J. Kelley, E. Khurana, J. Kidd, P. Lacroute, J. Lee, J. Leng, B. Li, H. Li, L. Li, Q. Long, G. Lunter, D. MacArthur, C. Melton, S. Montgomery, X. Mu, S. Musharoff, R. Nielsen, D. Presgraves, J. Pritchard, P. Sabeti, I. Shlyakhter, M. Snyder, C. Tyler-Smith, W. Wang, C. K. Zhang, Z. Zhang, H. Zhao, E. Eichler, M. Hurles, C. Lee, A. Abyzov, C. Alkan, G. del Angel, V. Bafna, E. Banks, A. Bashir, M. Batzer, E. Birney, B. Blackburne, T. Blackwell, M. Caccamo, N. Cardin, K. Cheetham, K. Chen, A. Chinwalla, L. Clarke, D. Conrad, D. Craig, J. Degenhardt, F. De La Vega, M. DePristo, L. Ding, M. Fromer, Y. Fu, E. Garrison, M. Gerstein, D. Greer, F. Grubert, G. Ha, M. Haimel, I. Hajirasouliha, B. Handsaker, R. Haraksingh, C. Hartl, N. Homer, F. Hormozdiari, S. Humphray, Z. Iqbal, S. Jones, M. Kaganovich, S. Kahn, P. Kersey, E. Khurana, J. Kidd, A. Ko, M. Konkel, J. Korbel, J. Korn, D. Kural, P. Lacroute, H. Lam, J. Lee, J. Leng, H. Li, Y. Li, G. Lunter, G. Marth, S. McCarroll, M. McLellan, R. Mills, V. Makarov, X. Mu, L. Murray, S. Nelson, J. Nemesh, Z. Ning, E. Parkhomenko, H. Peckham, A. Quinlan, S. Sajjadian, A. Scally, J. Sebat, K. Shakir, R. Shaw, M. Snyder, M. Snyder, C. Stewart, M. Stromberg, P. Sudmant, Y. Sun, D. Thomas, A. Urban, J. Wallis, J. Walker, K. Walter, A. Ward, J. Wu, C. Xiao, J. Xing, K. Ye, S. Yoon, D. Zerbino, Y. Zhang, Z. Zhang, H. Z. Bradley, R. Gibbs, G. Marth, G. Abecasis, K. Albers, M. Bainbridge, S. Balasubramaniam, S. Balasubramanian, T. Blackwell, C. Bustamanter, D. Challis, K. Cibulskis, A. Clark, L. Clarke, C. Coarfa, A. Coffrey, M. DePristo, M. Dermitzakis, D. Dooling, R. Durbin, T. Fennell, P. Flicek, L. Fulton, R. Fulton, S. Gabriel, K. Garimella, E. Garrison, R. Gibbs, S. Gravel, C. Hartl, A. Indap, Z. Iqbal, S. Katzman, T. Keane, D. Koboldt, W. F. Leong, D. MacArthur, E. Mardis, G. McVean, D. Muzny, E. Nickerson, A. Palotie, J. Paschall, J. Reid, A. Quinlan, A. Sabo, E. Shefler, S. Sherry, M. Shumway, R. Smith, C. Sougnez, J. Stalker, P. Sudmant, C. Tyler-Smith, Y. Wang, Y. Wang, J. Wilkinson, C. Xiao, F. Yu, H. Z. Bradley, L. Brooks, A. Felsenfeld, J. McEwen, N. Clemm, A. Duncanson, M. Dunn, E. Eichler, M. Guyer, M. Hurles, F. Hyland, R. Jamieson, N. Kalin, L. Kang, F. Laplace, C. Lee, Y. Liu, A. Michelson, J. Peterson, Z. Ren, and J. Wang. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug 2011. 61, 207

H. Daoud, H. Suhail, A. Szuto, W. Camu, F. Salachas, V. Meininger, J. P. Bouchard, N. Dupre, P. A. Dion, and G. A. Rouleau. UBQLN2 mutations are rare in French and French-Canadian amyotrophic lateral sclerosis. *Neurobiol. Aging*, 33(9):1–2230, Sep 2012. 23, 24

M. DeJesus-Hernandez, J. Kocerha, N. Finch, R. Crook, M. Baker, P. Desaro, A. Johnston, N. Rutherford, A. Wojtas, K. Kennelly, Z. K. Wszolek, N. Graff-Radford, K. Boylan, and R. Rademakers. De novo truncating FUS gene mutation as a cause of sporadic amyotrophic lateral sclerosis. *Hum. Mutat.*, 31(5):E1377–1389, May 2010. 18

M. DeJesus-Hernandez, I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker, N. J. Rutherford, A. M. Nicholson, N. A. Finch, H. Flynn, J. Adamson, N. Kouri, A. Wojtas, P. Sengdy, G. Y. Hsiung, A. Karydas, W. W. Seeley, K. A. Josephs, G. Coppola, D. H. Geschwind, Z. K. Wszolek, H. Feldman, D. S. Knopman, R. C. Petersen, B. L. Miller, D. W. Dickson, K. B. Boylan, N. R. Graff-Radford, and R. Rademakers. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, 72(2):245–256, Oct 2011. 11, 12, 20, 21, 22, 38, 271

H. Deng, K. Gao, and J. Jankovic. The role of FUS gene variants in neurodegenerative diseases. *Nat Rev Neurol*, 10(6):337–348, Jun 2014. 17, 18, 19

H. X. Deng, Y. Shi, Y. Furukawa, H. Zhai, R. Fu, E. Liu, G. H. Gorrie, M. S. Khan, W. Y. Hung, E. H. Bigio, T. Lukas, M. C. Dal Canto, T. V. O'Halloran, and T. Siddique. Conversion to the amyotrophic lateral sclerosis phenotype is associated with intermolecular linked insoluble aggregates of SOD1 in mitochondria. *Proc. Natl. Acad. Sci. U.S.A.*, 103(18):7142–7147, May 2006. 14

H. X. Deng, W. Chen, S. T. Hong, K. M. Boycott, G. H. Gorrie, N. Siddique, Y. Yang, F. Fecto, Y. Shi, H. Zhai, H. Jiang, M. Hirano, E. Rampersaud, G. H. Jansen, S. Donkervoort, E. H. Bigio, B. R. Brooks, K. Ajroud, R. L. Sufit, J. L. Haines, E. Mugnaini, M. A. Pericak-Vance, and T. Siddique. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature*, 477 (7363):211–215, Sep 2011. 8, 10, 12, 22, 23, 24, 34, 148, 167, 168

Y. D. Dewoody and J. A. Dewoody. On the estimation of genome-wide heterozygosity using molecular markers. *J. Hered.*, 96(2):85–88, 2005. 29

M. Di Salvio, V. Piccinni, V. Gerbino, F. Mantoni, S. Camerini, J. Lenzi, A. Rosa, L. Chellini, F. Loreni, M. T. Carri, I. Bozzoni, M. Cozzolino, and G. Cestra. Pur-alpha functionally interacts with FUS carrying ALS-associated mutations. *Cell Death Dis*, 6:e1943, Oct 2015. 154

P. A. Dion, H. Daoud, and G. A. Rouleau. Genetics of motor neuron disorders: new insights into pathogenic mechanisms. *Nat. Rev. Genet.*, 10(11):769–782, Nov 2009. 13

C. Dobson-Stone, A. D. Shaw, M. Hallupp, L. Bartley, H. McCann, W. S. Brooks, C. T. Loy, P. R. Schofield, K. A. Mather, N. A. Kochan, P. S. Sachdev, G. M. Halliday, O. Piguet, J. R. Hodges, and J. B. Kwok. Is CHCHD10 Pro34Ser pathogenic for frontotemporal dementia and amyotrophic lateral sclerosis? *Brain*, 138(Pt 10):e385, Oct 2015. 107

O. Dols-Icardo, A. Garcia-Redondo, R. Rojas-Garcia, R. Sanchez-Valle, A. Noguera, E. Gomez-Tortosa, P. Pastor, I. Hernandez, J. Esteban-Perez, M. Suarez-Calvet, S. Anton-Aguirre, G. Amer, S. Ortega-Cubero, R. Blesa, J. Fortea, D. Alcolea, A. Capdevila, A. Antonell, A. Llado, J. L. Munoz-Blanco, J. S. Mora, L. Galan-Davila, F. J. Rodriguez De Rivera, A. Lleo, and J. Clarimon. Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia. *Hum. Mol. Genet.*, 23(3):749–754, Feb 2014. 20, 21

O. Dols-Icardo, I. Nebot, A. Gorostidi, S. Ortega-Cubero, I. Hernandez, R. Rojas-Garcia, A. Garcia-Redondo, M. Povedano, A. Llado, V. Alvarez, P. Sanchez-Juan, J. Pardo, I. Jerico, J. Vazquez-Costa, T. Sevilla, F. Cardona, B. Indakoechea, F. Moreno, R. Fernandez-Torron, L. Munoz-Llahuna, S. Moreno-Grau, M. Rosende-Roca, A. Vela, J. L. Munoz-Blanco, O. Combarros, E. Coto, D. Alcolea, J. Fortea, A. Lleo, R. Sanchez-Valle, J. Esteban-Perez, A. Ruiz, P. Pastor, A. Lopez De Munain, J. Perez-Tur, and J. Clarimon. Analysis of the CHCHD10 gene in patients with frontotemporal dementia and amyotrophic lateral sclerosis from Spain. *Brain*, 138(Pt 12):e400, Dec 2015. 106, 143

O. Dols-Icardo, A. Garcia-Redondo, R. Rojas-Garcia, D. Borrego-Hernandez, I. Illan-Gala, J. L. Munoz-Blanco, A. Rabano, L. Cervera-Carles, A. Juarez-Rufian, N. Spataro, N. De Luna, L. Galan, E. Cortes-Vicente, J. Fortea, R. Blesa, O. Grau-Rivera, A. Lleo, J. Esteban-Perez, E. Gelpi, and J. Clarimon. Analysis of known amyotrophic lateral sclerosis and frontotemporal dementia genes reveals a substantial genetic burden in patients manifesting both diseases not carrying the C9orf72 expansion mutation. *J. Neurol. Neurosurg. Psychiatry*, 89(2):162–168, Feb 2018. 28

C. J. Donnelly, P. W. Zhang, J. T. Pham, A. R. Haeusler, A. R. Heusler, N. A. Mistry, S. Vidensky, E. L. Daley, E. M. Poth, B. Hoover, D. M. Fines, N. Maragakis, P. J. Tienari, L. Petrucelli, B. J. Traynor, J. Wang, F. Rigo, C. F. Bennett, S. Blackshaw, R. Sattler, and J. D. Rothstein. RNA toxicity from the ALS/FTD C9ORF72

expansion is mitigated by antisense intervention. *Neuron*, 80(2):415–428, Oct 2013. 7

D. Dormann, R. Rodde, D. Edbauer, E. Bentmann, I. Fischer, A. Hruscha, M. E. Than, I. R. Mackenzie, A. Capell, B. Schmid, M. Neumann, and C. Haass. ALS-associated fused in sarcoma (FUS) mutations disrupt Transportin-mediated nuclear import. *EMBO J.*, 29(16):2841–2857, Aug 2010. 19

M. Dowle and A. Srinivasan. *data.table: Extension of data.frame*, 2018. URL https://CRAN.R-project.org/package=data.table. 65

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca,

B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed,

J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, Sep 2012. 299

M. Elamin, P. Bede, S. Byrne, N. Jordan, L. Gallagher, B. Wynne, C. O'Brien, J. Phukan, C. Lynch, N. Pender, and O. Hardiman. Cognitive changes predict functional decline in ALS: a population-based longitudinal study. *Neurology*, 80(17): 1590–1597, Apr 2013. 5

A. C. Elden, H. J. Kim, M. P. Hart, A. S. Chen-Plotkin, B. S. Johnson, X. Fang, M. Armakola, F. Geser, R. Greene, M. M. Lu, A. Padmanabhan, D. Clay-Falcone,

L. McCluskey, L. Elman, D. Juhr, P. J. Gruber, U. Rub, G. Auburger, J. Q. Trojanowski, V. M. Lee, V. M. Van Deerlin, N. M. Bonini, and A. D. Gitler. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, 466(7310):1069–1075, Aug 2010. 9, 12, 27, 271, 283, 305

A. C. Fahed, B. McDonough, C. M. Gouvion, K. L. Newell, L. S. Dure, M. Bebin, A. G. Bick, J. G. Seidman, D. H. Harter, and C. E. Seidman. UBQLN2 mutation causing heterogeneous X-linked dominant neurodegeneration. *Ann. Neurol.*, 75(5): 793–798, May 2014. 23

M. A. Farg, V. Sundaramoorthy, J. M. Sultana, S. Yang, R. A. Atkinson, V. Levina, M. A. Halloran, P. A. Gleeson, I. P. Blair, K. Y. Soo, A. E. King, and J. D. Atkin. C9ORF72, implicated in amytrophic lateral sclerosis and frontotemporal dementia, regulates endosomal trafficking. *Hum. Mol. Genet.*, 23(13):3579–3595, Jul 2014. 21

K. D. Farwell, L. Shahmirzadi, D. El-Khechen, Z. Powis, E. C. Chao, B. Tippin Davis, R. M. Baxter, W. Zeng, C. Mroske, M. C. Parra, S. K. Gandomi, I. Lu, X. Li, H. Lu, H. M. Lu, D. Salvador, D. Ruble, M. Lao, S. Fischbach, J. Wen, S. Lee, A. Elliott, C. L. Dunlop, and S. Tang. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. *Genet. Med.*, 17(7):578–586, Jul 2015. 32

F. Fecto, J. Yan, S. P. Vemula, E. Liu, Y. Yang, W. Chen, J. G. Zheng, Y. Shi, N. Siddique, H. Arrat, S. Donkervoort, S. Ajroud-Driss, R. L. Sufit, S. L. Heller, H. X. Deng, and T. Siddique. SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis. *Arch. Neurol.*, 68(11):1440–1446, Nov 2011. 10, 12, 25, 148

R. Feil and M. F. Fraga. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.*, 13(2):97–109, Feb 2011. 283

E. Feneberg, E. Gray, O. Ansorge, K. Talbot, and M. R. Turner. Towards a TDP-43-Based Biomarker for ALS and FTLD. *Mol. Neurobiol.*, Feb 2018. 16, 17

A. Fernandez-Marmiesse, S. Gouveia, and M. L. Couce. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Curr. Med. Chem.*, 25 (3):404–432, Jan 2018. 31, 33, 289, 293, 300

A. Fernandez-Medarde, A. Porteros, J. de las Rivas, A. Nunez, J. J. Fuster, and E. Santos. Laser microdissection and microarray analysis of the hippocampus of Ras-GRF1

knockout mice reveals gene expression changes affecting signal transduction pathways related to memory and learning. *Neuroscience*, 146(1):272–285, Apr 2007. 217

M. A. Field, V. Cho, T. D. Andrews, and C. C. Goodnow. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. *PLoS ONE*, 10(11):e0143199, 2015. 270

D. A. Figlewicz, A. Krizus, M. G. Martinoli, V. Meininger, M. Dib, G. A. Rouleau, and J. P. Julien. Variants of the heavy neurofilament subunit are associated with the development of amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 3(10):1757–1761, Oct 1994. 12

C. Figueroa-Romero, J. Hur, D. E. Bender, C. E. Delaney, M. D. Cataldo, A. L. Smith, R. Yung, D. M. Ruden, B. C. Callaghan, and E. L. Feldman. Identification of epigenetically altered genes in sporadic amyotrophic lateral sclerosis. *PLoS ONE*, 7(12):e52672, 2012. 272, 284

D. Finley. Recognition and processing of ubiquitin-protein conjugates by the proteasome. *Annu. Rev. Biochem.*, 78:477–513, 2009. 10

I. Fogh, K. Lin, C. Tiloca, J. Rooney, C. Gellera, F. P. Diekstra, A. Ratti, A. Shatunov, M. A. van Es, P. Proitsi, A. Jones, W. Sproviero, A. Chio, R. L. McLaughlin, G. Soraru, L. Corrado, D. Stahl, R. Del Bo, C. Cereda, B. Castellotti, J. D. Glass, S. Newhouse, R. Dobson, B. N. Smith, S. Topp, W. van Rheenen, V. Meininger, J. Melki, K. E. Morrison, P. J. Shaw, P. N. Leigh, P. M. Andersen, G. P. Comi, N. Ticozzi, L. Mazzini, S. D'Alfonso, B. J. Traynor, P. Van Damme, W. Robberecht, R. H. Brown, J. E. Landers, O. Hardiman, C. M. Lewis, L. H. van den Berg, C. E. Shaw, J. H. Veldink, V. Silani, A. Al-Chalabi, and J. Powell. Association of a Locus in the CAMTA1 Gene With Survival in Patients With Sporadic Amyotrophic Lateral Sclerosis. *JAMA Neurol*, 73(7):812–820, Jul 2016. 154

M. S. Forman, J. Q. Trojanowski, and V. M. Lee. Neurodegenerative diseases: a decade of discoveries paves the way for therapeutic breakthroughs. *Nat. Med.*, 10 (10):1055–1063, Oct 2004. 7

B. D. Freibaum, Y. Lu, R. Lopez-Gonzalez, N. C. Kim, S. Almeida, K. H. Lee, N. Badders, M. Valentine, B. L. Miller, P. C. Wong, L. Petrucelli, H. J. Kim, F. B. Gao, and J. P. Taylor. GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. *Nature*, 525(7567):129–133, Sep 2015. 22

A. Freischmidt, T. Wieland, B. Richter, W. Ruf, V. Schaeffer, K. Muller, N. Marroquin, F. Nordin, A. Hubers, P. Weydt, S. Pinto, R. Press, S. Millecamps, N. Molko, E. Bernard, C. Desnuelle, M. H. Soriani, J. Dorst, E. Graf, U. Nordstrom, M. S. Feiler, S. Putz, T. M. Boeckers, T. Meyer, A. S. Winkler, J. Winkelman, M. de Carvalho, D. R. Thal, M. Otto, T. Brannstrom, A. E. Volk, P. Kursula, K. M. Danzer, P. Lichtner, I. Dikic, T. Meitinger, A. C. Ludolph, T. M. Strom, P. M. Andersen, and J. H. Weishaupt. Haploinsufficiency of TBK1 causes familial ALS and frontotemporal dementia. *Nat. Neurosci.*, 18(5):631–636, May 2015. 12, 25, 35, 282

I. Fridovich. Superoxide radical and superoxide dismutases. *Annu. Rev. Biochem.*, 64: 97–112, 1995. 14

M. Fromer, J. L. Moran, K. Chambert, E. Banks, S. E. Bergen, D. M. Ruderfer, R. E. Handsaker, S. A. McCarroll, M. C. O'Donovan, M. J. Owen, G. Kirov, P. F. Sullivan, C. M. Hultman, P. Sklar, and S. M. Purcell. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*, 91(4):597–607, Oct 2012. 305

K. Frousios, C. S. Iliopoulos, T. Schlitt, and M. A. Simpson. Predicting the functional consequences of non-synonymous DNA sequence variants–evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*, 102(4):223–228, Oct 2013. 297

H. Furukawa, S. Oka, T. Matsui, A. Hashimoto, Y. Arinuma, A. Komiya, N. Fukui, N. Tsuchiya, and S. Tohma. Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. *Hum. Immunol.*, 74(2):170–175, Feb 2013. 267, 279

B. Gaastra, A. Shatunov, S. Pulit, A. R. Jones, W. Sproviero, A. Gillett, Z. Chen, J. Kirby, I. Fogh, J. F. Powell, P. N. Leigh, K. E. Morrison, P. J. Shaw, C. E. Shaw, L. H. van den Berg, J. H. Veldink, C. M. Lewis, and A. Al-Chalabi. Rare genetic variation in UNC13A may modify survival in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener*, 17(7-8):593–599, 2016. 27

W. A. Gahl, T. C. Markello, C. Toro, K. F. Fajardo, M. Sincan, F. Gill, H. Carlson-Donohoe, A. Gropman, T. M. Pierson, G. Golas, L. Wolfe, C. Groden, R. Godfrey, M. Nehrebecky, C. Wahl, D. M. Landis, S. Yang, A. Madeo, J. C. Mullikin, C. F. Boerkoel, C. J. Tifft, and D. Adams. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet. Med.*, 14(1):51–59, Jan 2012. 32

J. Gamez, M. Corbera-Bellalta, G. Nogales, N. Raguer, E. Garcia-Arumi, M. Badia-Canto, E. Llado-Carbo, and J. Alvarez-Sabin. Mutational analysis of the Cu/Zn superoxide dismutase gene in a Catalan ALS population: should all sporadic ALS cases also be screened for SOD1? *J. Neurol. Sci.*, 247(1):21–28, Aug 2006. 26

S. Gazal, S. Gosset, E. Verdura, F. Bergametti, S. Guey, M. C. Babron, and E. Tournier-Lasserve. Can whole-exome sequencing data be used for linkage analysis? *Eur. J. Hum. Genet.*, 24(4):581–586, Apr 2016. 38

C. Gellera, C. Tiloca, R. Del Bo, L. Corrado, V. Pensato, J. Agostini, C. Cereda, A. Ratti, B. Castellotti, S. Corti, A. Bagarotti, A. Cagnin, P. Milani, C. Gabelli, G. Riboldi, L. Mazzini, G. Soraru, S. D'Alfonso, F. Taroni, G. P. Comi, N. Ticozzi, and V. Silani. Ubiquilin 2 mutations in Italian patients with amyotrophic lateral sclerosis and frontotemporal dementia. *J. Neurol. Neurosurg. Psychiatry*, 84(2):183–187, Feb 2013. 23, 24

T. F. Gendron, K. F. Bieniek, Y. J. Zhang, K. Jansen-West, P. E. Ash, T. Caulfield, L. Daughrity, J. H. Dunmore, M. Castanedes-Casey, J. Chew, D. M. Cosio, M. van Blitterswijk, W. C. Lee, R. Rademakers, K. B. Boylan, D. W. Dickson, and L. Petrucelli. Antisense transcripts of the expanded C9ORF72 hexanucleotide repeat form nuclear RNA foci and undergo repeat-associated non-ATG translation in c9FTD/ALS. *Acta Neuropathol.*, 126(6):829–844, Dec 2013. 22

T. F. Gendron, V. V. Belzil, Y. J. Zhang, and L. Petrucelli. Mechanisms of toxicity in C9FTLD/ALS. *Acta Neuropathol.*, 127(3):359–376, Mar 2014. 21

I. Gijselinck, S. Engelborghs, G. Maes, I. Cuijt, K. Peeters, M. Mattheijssens, G. Joris, P. Cras, J. J. Martin, P. P. De Deyn, S. Kumar-Singh, C. Van Broeckhoven, and M. Cruts. Identification of 2 Loci at chromosomes 9 and 14 in a multiplex family with frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Arch. Neurol.*, 67(5):606–616, May 2010. 19

I. Gijselinck, T. Van Langenhove, J. van der Zee, K. Sleegers, S. Philtjens, G. Kleinberger, J. Janssens, K. Bettens, C. Van Cauwenberghe, S. Pereson, S. Engelborghs, A. Sieben, P. De Jonghe, R. Vandenberghe, P. Santens, J. De Bleecker, G. Maes, V. Baumer, L. Dillen, G. Joris, I. Cuijt, E. Corsmit, E. Elinck, J. Van Dongen, S. Vermeulen, M. Van den Broeck, C. Vaerenberg, M. Mattheijssens, K. Peeters, W. Robberecht, P. Cras, J. J. Martin, P. P. De Deyn, M. Cruts, and C. Van Broeckhoven. A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders

of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum: a gene identification study. *Lancet Neurol*, 11(1):54–65, Jan 2012. 21

I. Gijselinck, S. Van Mossevelde, J. van der Zee, A. Sieben, S. Engelborghs, J. De Bleecker, A. Ivanoiu, O. Deryck, D. Edbauer, M. Zhang, B. Heeman, V. Baumer, M. Van den Broeck, M. Mattheijssens, K. Peeters, E. Rogaeva, P. De Jonghe, P. Cras, J. J. Martin, P. P. de Deyn, M. Cruts, and C. Van Broeckhoven. The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol. Psychiatry*, 21(8): 1112–1124, 08 2016. 21

C. Gilissen, H. H. Arts, A. Hoischen, L. Spruijt, D. A. Mans, P. Arts, B. van Lier, M. Steehouwer, J. van Reeuwijk, S. G. Kant, R. Roepman, N. V. Knoers, J. A. Veltman, and H. G. Brunner. Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.*, 87(3):418–423, Sep 2010. 32

C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, R. Leach, R. Klein, R. Tearle, T. Bo, R. Pfundt, H. G. Yntema, B. B. de Vries, T. Kleefstra, H. G. Brunner, L. E. Vissers, and J. A. Veltman. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509):344–347, Jul 2014. 33, 286

D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.*, 14(7):460–470, Jul 2013. 296

M. J. Greenway, M. D. Alexander, S. Ennis, B. J. Traynor, B. Corr, E. Frost, A. Green, and O. Hardiman. A novel candidate region for ALS on chromosome 14q11.2. *Neurology*, 63(10):1936–1938, Nov 2004. 25

M. J. Greenway, P. M. Andersen, C. Russ, S. Ennis, S. Cashman, C. Donaghy, V. Patterson, R. Swingler, D. Kieran, J. Prehn, K. E. Morrison, A. Green, K. R. Acharya, R. H. Brown, and O. Hardiman. ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat. Genet.*, 38(4):411–413, Apr 2006. 9, 12, 25

E. J. Groen, K. Fumoto, A. M. Blokhuis, J. Engelen-Lee, Y. Zhou, D. M. van den Heuvel, M. Koppers, F. van Diggelen, J. van Heest, J. A. Demmers, J. Kirby, P. J. Shaw, E. Aronica, W. G. Spliet, J. H. Veldink, L. H. van den Berg, and R. J. Pasterkamp. ALS-associated mutations in FUS disrupt the axonal distribution and function of SMN. *Hum. Mol. Genet.*, 22(18):3690–3704, Sep 2013. 19

F. Gros-Louis, R. Lariviere, G. Gowing, S. Laurent, W. Camu, J. P. Bouchard, V. Meininger, G. A. Rouleau, and J. P. Julien. A frameshift deletion in peripherin gene associated with amyotrophic lateral sclerosis. *J. Biol. Chem.*, 279(44): 45951–45956, Oct 2004. 12

X. Gu, Y. Chen, Q. Wei, B. Cao, R. Ou, X. Yuan, Y. Hou, L. Zhang, H. Liu, X. Chen, and H. F. Shang. Mutation screening of the TIA1 gene in Chinese patients with amyotrophic lateral sclerosis/frontotemporal dementia. *Neurobiol. Aging*, 68:1–161, Aug 2018. 161, 278

M. H. Guo, A. Dauber, M. F. Lippincott, Y. M. Chan, R. M. Salem, and J. N. Hirschhorn. Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *Am. J. Hum. Genet.*, 99(3):527–539, Sep 2016. 27

M. E. Gurney, H. Pu, A. Y. Chiu, M. C. Dal Canto, C. Y. Polchow, D. D. Alexander, J. Caliendo, A. Hentati, Y. W. Kwon, and H. X. Deng. Motor neuron degeneration in mice that express a human Cu,Zn superoxide dismutase mutation. *Science*, 264 (5166):1772–1775, Jun 1994. 14

S. Hadano, C. K. Hand, H. Osuga, Y. Yanagisawa, A. Otomo, R. S. Devon, N. Miyamoto, J. Showguchi-Miyata, Y. Okada, R. Singaraja, D. A. Figlewicz, T. Kwiatkowski, B. A. Hosler, T. Sagie, J. Skaug, J. Nasir, R. H. Brown, S. W. Scherer, G. A. Rouleau, M. R. Hayden, and J. E. Ikeda. A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat. Genet.*, 29(2):166–173, Oct 2001. 12, 25

S. P. Hagenaars, R. Radakovi, C. Crockford, C. Fawns-Ritchie, S. E. Harris, C. R. Gale, and I. J. Deary. Genetic risk for neurodegenerative disorders, and its overlap with cognitive ability and physical function. *PLoS ONE*, 13(6):e0198187, 2018. 281

A. E. Handel, G. C. Ebers, and S. V. Ramagopalan. Epigenetics: molecular mechanisms and implications for disease. *Trends Mol Med*, 16(1):7–16, Jan 2010. 283

Y. Hayashi, K. Homma, and H. Ichijo. SOD1 in neurotoxicity and its controversial roles in SOD1 mutation-negative ALS. *Adv Biol Regul*, 60:95–104, Jan 2016. 14

J. B. Hilton, S. W. Mercer, N. K. Lim, N. G. Faux, G. Buncic, J. S. Beckman, B. R. Roberts, P. S. Donnelly, A. R. White, and P. J. Crouch. CuII(atsm) improves the neurological phenotype and survival of SOD1G93A mice and selectively increases enzymatically active SOD1 in the spinal cord. *Sci Rep*, 7:42292, Feb 2017. 144

A. Hoischen, B. W. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, K. Devriendt, M. Z. Amorim, N. Revencu, A. Kidd, M. Barbosa, A. Turner, J. Smith, C. Oley, A. Henderson, I. M. Hayes, E. M. Thompson, H. G. Brunner, B. B. de Vries, and J. A. Veltman. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.*, 42 (6):483–485, Jun 2010. 32

E. P. Hong and J. W. Park. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*, 10(2):117–122, Jun 2012. 164

L. Hou, B. Jiao, T. Xiao, L. Zhou, Z. Zhou, J. Du, X. Yan, J. Wang, B. Tang, and L. Shen. Screening of SOD1, FUS and TARDBP genes in patients with amyotrophic lateral sclerosis in central-southern China. *Sci Rep*, 6:32478, Sep 2016. 13

H. Hu, J. C. Roach, H. Coon, S. L. Guthery, K. V. Voelkerding, R. L. Margraf, J. D. Durtschi, S. V. Tavtigian, S. Wu, W. Wu, P. Scheet, S. Wang, J. Xing, G. Glusman, R. Hubley, H. Li, V. Garg, B. Moore, L. Hood, D. J. Galas, D. Srivastava, M. G. Reese, L. B. Jorde, M. Yandell, and C. D. Huff. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature Biotechnol*, 32(7):663–669, 2014. 38, 278

H. Hu, S. A. Haas, J. Chelly, H. Van Esch, M. Raynaud, A. P. de Brouwer, S. Weinert, G. Froyen, S. G. Frints, F. Laumonnier, T. Zemojtel, M. I. Love, H. Richard, A. K. Emde, M. Bienek, C. Jensen, M. Hambrock, U. Fischer, C. Langnick, M. Feldkamp, W. Wissink-Lindhout, N. Lebrun, L. Castelnau, J. Rucci, R. Montjean, O. Dorseuil, P. Billuart, T. Stuhlmann, M. Shaw, M. A. Corbett, A. Gardner, S. Willis-Owen, C. Tan, K. L. Friend, S. Belet, K. E. van Roozendaal, M. Jimenez-Pocquet, M. P. Moizard, N. Ronce, R. Sun, S. O'Keeffe, R. Chenna, A. van Bommel, J. Goke, A. Hackett, M. Field, L. Christie, J. Boyle, E. Haan, J. Nelson, G. Turner, G. Baynam, G. Gillessen-Kaesbach, U. Muller, D. Steinberger, B. Budny, M. Badura-Stronka, A. Latos-Bielenska, L. B. Ousager, P. Wieacker, G. Rodriguez Criado, M. L. Bondeson, G. Anneren, A. Dufke, M. Cohen, L. Van Maldergem, C. Vincent-Delorme, B. Echenne, B. Simon-Bouy, T. Kleefstra, M. Willemsen, J. P. Fryns, K. Devriendt, R. Ullmann, M. Vingron, K. Wrogemann, T. F. Wienker, A. Tzschach, H. van Bokhoven, J. Gecz, T. J. Jentsch, W. Chen, H. H. Ropers, and V. M. Kalscheuer. X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol. Psychiatry*, 21(1):133–148, Jan 2016. 214

Y. F. Huang, B. Gulko, and A. Siepel. Fast, scalable prediction of deleterious noncoding

variants from functional and population genomic data. *Nat. Genet.*, 49(4):618–624, Apr 2017. 299

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. URL http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html. 63

M. H. Huisman, S. W. de Jong, P. T. van Doormaal, S. S. Weinreich, H. J. Schelhaas, A. J. van der Kooi, M. de Visser, J. H. Veldink, and L. H. van den Berg. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J. Neurol. Neurosurg. Psychiatry*, 82(10):1165–1170, Oct 2011. 3

S. Hwang, E. Kim, I. Lee, and E. M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*, 5:17875, Dec 2015. 270, 288, 289, 293

L. Ibanez, F. Farias, U. Dube, K. Mihindukulasuriya, and O. Harari. Polygenic Risk Scores in Neurodegenerative Diseases: a Review. *Curr. Genet. Med. Rep.*, 7:22–29, Sep 2019. 281

L. M. Igaz, L. K. Kwong, A. Chen-Plotkin, M. J. Winton, T. L. Unger, Y. Xu, M. Neumann, J. Q. Trojanowski, and V. M. Lee. Expression of TDP-43 C-terminal Fragments in Vitro Recapitulates Pathological Features of TDP-43 Proteinopathies. *J. Biol. Chem.*, 284(13):8516–8524, Mar 2009. 16

Illumina. *An introduction to Next-Generation Sequencing Technology*. Illumina Inc., 2017. 45, 48

C. Ingre, P. M. Roos, F. Piehl, F. Kamel, and F. Fang. Risk factors for amyotrophic lateral sclerosis. *Clin Epidemiol*, 7:181–193, 2015. 5, 6

B. Jiao, T. Xiao, L. Hou, X. Gu, Y. Zhou, L. Zhou, B. Tang, J. Xu, and L. Shen. High prevalence of CHCHD10 mutation in patients with frontotemporal dementia from China. *Brain*, 139(Pt 4):e21, Apr 2016. 106

J. O. Johnson, J. Mandrioli, M. Benatar, Y. Abramzon, V. M. Van Deerlin, J. Q. Trojanowski, J. R. Gibbs, M. Brunetti, S. Gronka, J. Wuu, J. Ding, L. McCluskey,

M. Martinez-Lage, D. Falcone, D. G. Hernandez, S. Arepalli, S. Chong, J. C. Schymick, J. Rothstein, F. Landi, Y. D. Wang, A. Calvo, G. Mora, M. Sabatelli, M. R. Monsurro, S. Battistini, F. Salvi, R. Spataro, P. Sola, G. Borghero, G. Galassi, S. W. Scholz, J. P. Taylor, G. Restagno, A. Chio, B. J. Traynor, F. Giannini, C. Ricci, C. Moglia, I. Ossola, A. Canosa, S. Gallo, G. Tedeschi, P. Sola, I. Bartolomei, K. Marinou, L. Papetti, A. Conte, M. Luigetti, V. La Bella, P. Paladino, C. Caponnetto, P. Volanti, M. G. Marrosu, and M. R. Murru. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron*, 68(5):857–864, Dec 2010. 10, 12, 25, 34

J. O. Johnson, S. M. Glynn, J. R. Gibbs, M. A. Nalls, M. Sabatelli, G. Restagno, V. E. Drory, A. Chio, E. Rogaeva, and B. J. Traynor. Mutations in the CHCHD10 gene are a common cause of familial amyotrophic lateral sclerosis. *Brain*, 137(Pt 12):e311, Dec 2014a. 106

J. O. Johnson, E. P. Pioro, A. Boehringer, R. Chia, H. Feit, A. E. Renton, H. A. Pliner, Y. Abramzon, G. Marangi, B. J. Winborn, J. R. Gibbs, M. A. Nalls, S. Morgan, M. Shoai, J. Hardy, A. Pittman, R. W. Orrell, A. Malaspina, K. C. Sidle, P. Fratta, M. B. Harms, R. H. Baloh, A. Pestronk, C. C. Weihl, E. Rogaeva, L. Zinman, V. E. Drory, G. Borghero, G. Mora, A. Calvo, J. D. Rothstein, C. Drepper, M. Sendtner, A. B. Singleton, J. P. Taylor, M. R. Cookson, G. Restagno, M. Sabatelli, R. Bowser, A. Chio, B. J. Traynor, C. Moglia, S. Cammarosano, A. Canosa, S. Gallo, M. Brunetti, I. Ossola, K. Marinou, L. Papetti, F. Pisano, G. L. Pinter, A. Conte, M. Luigetti, M. Zollino, S. Lattante, G. Marangi, V. La Bella, R. Spataro, T. Colletti, F. Giannini, S. Battistini, C. Ricci, C. Caponnetto, G. Mancardi, P. Mandich, F. Salvi, I. Bartolomei, J. Mandrioli, P. Sola, C. Lunetta, S. Penco, M. R. Monsurro, G. Tedeschi, F. L. Conforti, A. Gambardella, A. Quattrone, P. Volanti, G. Floris, A. Cannas, V. Piras, F. Marrosu, M. G. Marrosu, M. R. Murru, M. Pugliatti, L. D. Parish, A. Sotgiu, G. Solinas, L. Ulgheri, A. Ticca, I. Simone, G. Logroscino, and A. Pirisi. Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. *Nat. Neurosci.*, 17(5):664–666, May 2014b. 12, 25, 34

C. A. Johnston, B. R. Stanton, M. R. Turner, R. Gray, A. H. Blunt, D. Butt, M. A. Ampong, C. E. Shaw, P. N. Leigh, and A. Al-Chalabi. Amyotrophic lateral sclerosis in an urban setting: a population based study of inner city London. *J. Neurol.*, 253 (12):1642–1643, Dec 2006. 5

A. R. Jones, I. Woollacott, A. Shatunov, J. Cooper-Knock, V. Buchman, W. Sproviero, B. Smith, K. M. Scott, R. Balendra, O. Abel, P. McGuffin, C. M. Ellis, P. J. Shaw,

K. E. Morrison, A. Farmer, C. M. Lewis, P. N. Leigh, C. E. Shaw, J. F. Powell, and A. Al-Chalabi. Residual association at C9orf72 suggests an alternative amyotrophic lateral sclerosis-causing hexanucleotide repeat. *Neurobiol. Aging*, 34(9): 1–7, Sep 2013. 145

E. Kabashi, P. N. Valdmanis, P. Dion, D. Spiegelman, B. J. McConkey, C. Vande Velde, J. P. Bouchard, L. Lacomblez, K. Pochigaeva, F. Salachas, P. F. Pradat, W. Camu, V. Meininger, N. Dupre, and G. A. Rouleau. TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.*, 40(5):572–574, May 2008. 15

E. Kabashi, L. Lin, M. L. Tradewell, P. A. Dion, V. Bercier, P. Bourgouin, D. Rochefort, S. Bel Hadj, H. D. Durham, C. Vande Velde, G. A. Rouleau, and P. Drapeau. Gain and loss of function of ALS-related mutations of TARDBP (TDP-43) cause motor deficits in vivo. *Hum. Mol. Genet.*, 19(4):671–683, Feb 2010. 16

H. M. Kaneb, A. W. Folkmann, V. V. Belzil, L. E. Jao, C. S. Leblond, S. L. Girard, H. Daoud, A. Noreau, D. Rochefort, P. Hince, A. Szuto, A. Levert, S. Vidal, C. Andre-Guimont, W. Camu, J. P. Bouchard, N. Dupre, G. A. Rouleau, S. R. Wente, and P. A. Dion. Deleterious mutations in the essential mRNA metabolism factor, hGle1, in amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 24(5):1363–1373, Mar 2015. 12, 25, 148, 154

H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, and N. Sestan. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, Oct 2011. 55, 57, 195

B. A. Keller, K. Volkening, C. A. Droppelmann, L. C. Ang, R. Rademakers, and M. J. Strong. Co-aggregation of RNA binding proteins in ALS spinal motor neurons: evidence of a common pathogenic mechanism. *Acta Neuropathol.*, 124(5):733–747, Nov 2012. 15

K. P. Kenna, P. T. van Doormaal, A. M. Dekker, N. Ticozzi, B. J. Kenna, F. P. Diekstra, W. van Rheenen, K. R. van Eijk, A. R. Jones, P. Keagle, A. Shatunov, W. Sproviero, B. N. Smith, M. A. van Es, S. D. Topp, A. Kenna, J. W. Miller, C. Fallini, C. Tiloca, R. L. McLaughlin, C. Vance, C. Troakes, C. Colombrita,

G. Mora, A. Calvo, F. Verde, S. Al-Sarraj, A. King, D. Calini, J. de Belleroche, F. Baas, A. J. van der Kooi, M. de Visser, A. L. Ten Asbroek, P. C. Sapp, D. McKenna-Yasek, M. Polak, S. Asress, J. L. Munoz-Blanco, T. M. Strom, T. Meitinger, K. E. Morrison, G. Lauria, K. L. Williams, P. N. Leigh, G. A. Nicholson, I. P. Blair, C. S. Leblond, P. A. Dion, G. A. Rouleau, H. Pall, P. J. Shaw, M. R. Turner, K. Talbot, F. Taroni, K. B. Boylan, M. Van Blitterswijk, R. Rademakers, J. Esteban-Perez, A. Garcia-Redondo, P. Van Damme, W. Robberecht, A. Chio, C. Gellera, C. Drepper, M. Sendtner, A. Ratti, J. D. Glass, J. S. Mora, N. A. Basak, O. Hardiman, A. C. Ludolph, P. M. Andersen, J. H. Weishaupt, R. H. Brown, A. Al-Chalabi, V. Silani, C. E. Shaw, L. H. van den Berg, J. H. Veldink, J. E. Landers, S. D'Alfonso, L. Mazzini, G. P. Comi, R. Del Bo, M. Ceroni, S. Gagliardi, G. Querin, C. Bertolin, V. Pensato, B. Castellotti, S. Corti, C. Cereda, L. Corrado, and G. Soraru. NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, 48(9):1037–1042, 09 2016. 12, 25, 28, 35, 280, 282

M. J. Keogh, W. Wei, J. Aryaman, I. Wilson, K. Talbot, M. R. Turner, C. A. McKenzie, C. Troakes, J. Attems, C. Smith, S. Al Sarraj, C. M. Morris, O. Ansorge, S. Pickering-Brown, N. Jones, J. W. Ironside, and P. F. Chinnery. Oligogenic genetic variation of neurodegenerative disease genes in 980 postmortem human brains. *J. Neurol. Neurosurg. Psychiatry*, 89(8):813–816, Aug 2018. 279, 280

E. Kichkin, A. Visvanathan, F. J. Lovicu, D. Y. Shu, S. J. Das, S. W. Reddel, E. P. McCann, K. Y. Zhang, K. L. Williams, I. P. Blair, and W. D. Phillips. Postnatal Development of Spasticity Following Transgene Insertion in the Mouse IV Spectrin Gene (SPTBN4). *J Neuromuscul Dis*, 4(2):159–164, 2017. 154

M. C. Kiernan. Motor neuron disease in 2017: Progress towards therapy in motor neuron disease. *Nat Rev Neurol*, 14(2):65–66, Feb 2018. 7

M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing. Amyotrophic lateral sclerosis. *Lancet*, 377(9769):942–955, Mar 2011. 3, 4

H. J. Kim, N. C. Kim, Y. D. Wang, E. A. Scarborough, J. Moore, Z. Diaz, K. S. MacLea, B. Freibaum, S. Li, A. Molliex, A. P. Kanagaraj, R. Carter, K. B. Boylan, A. M. Wojtas, R. Rademakers, J. L. Pinkus, S. A. Greenberg, J. Q. Trojanowski, B. J. Traynor, B. N. Smith, S. Topp, A. S. Gkazi, J. Miller, C. E. Shaw, M. Kottlors, J. Kirschner, A. Pestronk, Y. R. Li, A. F. Ford, A. D. Gitler, M. Benatar, O. D. King, V. E. Kimonis, E. D. Ross, C. C. Weihl, J. Shorter, and J. P. Taylor. Mutations in

prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*, 495(7442):467–473, Mar 2013. 9, 12, 16, 19, 25, 34, 154

H. J. Kim, M. J. Kwon, W. J. Choi, K. W. Oh, S. I. Oh, C. S. Ki, and S. H. Kim. Mutations in UBQLN2 and SIGMAR1 genes are rare in Korean patients with amyotrophic lateral sclerosis. *Neurobiol. Aging*, 35(8):7–8, Aug 2014. 23

Y. Kino, C. Washizu, M. Kurosawa, M. Yamada, H. Miyazaki, T. Akagi, T. Hashikawa, H. Doi, T. Takumi, G. G. Hicks, N. Hattori, T. Shimogori, and N. Nukina. FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. *Acta Neuropathol Commun*, 3:24, Apr 2015. 18

M. Kircher, P. Heyn, and J. Kelso. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*, 12:382, Jul 2011. 292

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–315, Mar 2014. 57, 299

T. Kitada, S. Asakawa, N. Hattori, H. Matsumine, Y. Yamamura, S. Minoshima, M. Yokochi, Y. Mizuno, and N. Shimizu. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676):605–608, Apr 1998. 154

M. F. Kleijnen, A. H. Shih, P. Zhou, S. Kumar, R. E. Soccio, N. L. Kedersha, G. Gill, and P. M. Howley. The hPLIC proteins may provide a link between the ubiquitination machinery and the proteasome. *Mol. Cell*, 6(2):409–419, Aug 2000. 24

S. Kondo, B. C. Schutte, R. J. Richardson, B. C. Bjork, A. S. Knight, Y. Watanabe, E. Howard, R. L. de Lima, S. Daack-Hirsch, A. Sander, D. M. McDonald-McGinn, E. H. Zackai, E. J. Lammer, A. S. Aylsworth, H. H. Ardinger, A. C. Lidral, B. R. Pober, L. Moreno, M. Arcos-Burgos, C. Valencia, C. Houdayer, M. Bahuau, D. Moretti-Ferreira, A. Richieri-Costa, M. J. Dixon, and J. C. Murray. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat. Genet.*, 32 (2):285–289, Oct 2002. 222

N. J. Kramer, M. S. Haney, D. W. Morgens, A. Jovicic, J. Couthouis, A. Li, J. Ousey, R. Ma, G. Bieri, C. K. Tsui, Y. Shi, N. T. Hertz, M. Tessier-Lavigne, J. K. Ichida, M. C. Bassik, and A. D. Gitler. CRISPR-Cas9 screens in human cells and primary neurons identify modifiers of C9ORF72 dipeptide-repeat-protein toxicity. *Nat. Genet.*, 50(4):603–612, Apr 2018. 9

L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, 58 (6):1347–1363, Jun 1996. 30

P. Kuhnlein, H. Jung, M. Farkas, S. Keskitalo, B. Ineichen, I. Jelcic, J. Petersen, A. Semmler, M. Weller, A. C. Ludolph, and M. Linnebank. The thermolabile variant of 5,10-methylenetetrahydrofolate reductase is a possible risk factor for amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 12(2):136–139, Mar 2011. 154

P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7):1073–1081, 2009. 57

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, N. Abdennur, M. Adli, M. Akerman, L. Barrera, J. Antosiewicz-Bourget, T. Ballinger,

M. J. Barnes, D. Bates, R. J. Bell, D. A. Bennett, K. Bianco, C. Bock, P. Boyle, J. Brinchmann, P. Caballero-Campo, R. Camahort, M. J. Carrasco-Alfonso, T. Charnecki, H. Chen, Z. Chen, J. B. Cheng, S. Cho, A. Chu, W. Y. Chung, C. Cowan, Q. Athena Deng, V. Deshpande, M. Diegel, B. Ding, T. Durham, L. Echipare, L. Edsall, D. Flowers, O. Genbacev-Krtolica, C. Gifford, S. Gillespie, E. Giste, I. A. Glass, A. Gnirke, M. Gormley, H. Gu, J. Gu, D. A. Hafler, M. J. Hangauer, M. Hariharan, M. Hatan, E. Haugen, Y. He, S. Heimfeld, S. Herlofsen, Z. Hou, R. Humbert, R. Issner, A. R. Jackson, H. Jia, P. Jiang, A. K. Johnson, T. Kadlecek, B. Kamoh, M. Kapidzic, J. Kent, A. Kim, M. Kleinewietfeld, S. Klugman, J. Krishnan, S. Kuan, T. Kutyavin, A. Y. Lee, K. Lee, J. Li, N. Li, Y. Li, K. L. Ligon, S. Lin, Y. Lin, J. Liu, Y. Liu, C. J. Luckey, Y. P. Ma, C. Maire, A. Marson, J. S. Mattick, M. Mayo, M. McMaster, H. Metsky, T. Mikkelsen, D. Miller, M. Miri, E. Mukamel, R. P. Nagarajan, F. Neri, J. Nery, T. Nguyen, H. O'Geen, S. Paithankar, T. Papayannopoulou, M. Pelizzola, P. Plettner, N. E. Propson, S. Raghuraman, B. J. Raney, A. Raubitschek, A. P. Reynolds, H. Richards, K. Riehle, P. Rinaudo, J. F. Robinson, N. B. Rockweiler, E. Rosen, E. Rynes, J. Schein, R. Sears, T. Sejnowski, A. Shafer, L. Shen, R. Shoemaker, M. Sigaroudinia, I. Slukvin, S. Stehling-Sun, R. Stewart, S. L. Subramanian, K. Suknuntha, S. Swanson, S. Tian, H. Tilden, L. Tsai, M. Urich, I. Vaughn, J. Vierstra, S. Vong, U. Wagner, H. Wang, T. Wang, Y. Wang, A. Weiss, H. Whitton, A. Wildberg, H. Witt, K. J. Won, M. Xie, X. Xing, I. Xu, Z. Xuan, Z. Ye, C. A. Yen, P. Yu, X. Zhang, X. Zhang, J. Zhao, Y. Zhou, J. Zhu, Y. Zhu, S. Ziegler, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb 2015. 299

D. Kurzwelly, S. Kruger, S. Biskup, and M. T. Heneka. A distinct clinical phenotype in a German kindred with motor neuron disease carrying a CHCHD10 mutation. *Brain*, 138(Pt 9):e376, Sep 2015. 106, 143

T. J. Kwiatkowski, D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, J. Gilchrist, E. J. Kasarskis, T. Munsat, P. Valdmanis, G. A. Rouleau, B. A. Hosler, P. Cortelli, P. J. de Jong, Y. Yoshinaga, J. L. Haines, M. A. Pericak-Vance, J. Yan, N. Ticozzi, T. Siddique, D. McKenna-Yasek, P. C. Sapp,

H. R. Horvitz, J. E. Landers, and R. H. Brown. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science*, 323(5918): 1205–1208, Feb 2009. 11, 12, 17, 18, 19

C. T. Kwok, A. G. Morris, J. Frampton, B. Smith, C. E. Shaw, and J. de Belleroche. Association studies indicate that protein disulfide isomerase is a risk factor in amyotrophic lateral sclerosis. *Free Radic. Biol. Med.*, 58:81–86, May 2013. 12

M. J. Kwon, W. Baek, C. S. Ki, H. Y. Kim, S. H. Koh, J. W. Kim, and S. H. Kim. Screening of the SOD1, FUS, TARDBP, ANG, and OPTN mutations in Korean patients with familial and sporadic ALS. *Neurobiol. Aging*, 33(5):17–23, May 2012. 13

A. R. La Spada and J. P. Taylor. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.*, 11(4):247–258, Apr 2010. 22

A. R. La Spada, E. M. Wilson, D. B. Lubahn, A. E. Harding, and K. H. Fischbeck. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, 352(6330):77–79, Jul 1991. 283, 305

H. Laaksovirta, T. Peuralinna, J. C. Schymick, S. W. Scholz, S. L. Lai, L. Myllykangas, R. Sulkava, L. Jansson, D. G. Hernandez, J. R. Gibbs, M. A. Nalls, D. Heckerman, P. J. Tienari, and B. J. Traynor. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol*, 9(10):978–985, Oct 2010. 19, 27, 281

E. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, 11(3):241–247, Nov 1995. 30

J. E. Landers, J. Melki, V. Meininger, J. D. Glass, L. H. van den Berg, M. A. van Es, P. C. Sapp, P. W. van Vught, D. M. McKenna-Yasek, H. M. Blauw, T. J. Cho, M. Polak, L. Shi, A. M. Wills, W. J. Broom, N. Ticozzi, V. Silani, A. Ozoguz, I. Rodriguez-Leyva, J. H. Veldink, A. J. Ivinson, C. G. Saris, B. A. Hosler, A. Barnes-Nessa, N. Couture, J. H. Wokke, T. J. Kwiatkowski, R. A. Ophoff, S. Cronin, O. Hardiman, F. P. Diekstra, P. N. Leigh, C. E. Shaw, C. L. Simpson, V. K. Hansen, J. F. Powell, P. Corcia, F. Salachas, S. Heath, P. Galan, F. Georges, H. R. Horvitz, M. Lathrop, S. Purcell, A. Al-Chalabi, and R. H. Brown. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U.S.A.*, 106(22):9004–9009, Jun 2009. 27, 154

S. Lattante, G. A. Rouleau, and E. Kabashi. TARDBP and FUS mutations associated with amyotrophic lateral sclerosis: summary and update. *Hum. Mutat.*, 34(6):812–826, Jun 2013. 15, 16, 17, 18

B. Laube, H. Hirai, M. Sturgess, H. Betz, and J. Kuhse. Molecular determinants of agonist discrimination by NMDA receptor subunits: analysis of the glutamate binding site on the NR2B subunit. *Neuron*, 18(3):493–503, Mar 1997. 215

S. Laurie, M. Fernandez-Callejo, S. Marco-Sola, J. R. Trotta, J. Camps, A. Chacon, A. Espinosa, M. Gut, I. Gut, S. Heath, and S. Beltran. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Hum. Mutat.*, 37(12):1263–1271, 12 2016. 269, 286, 287, 288, 290, 291, 293, 295

R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, 15(6):R84, Jun 2014. 305

I. Le Ber, A. Camuzat, E. Berger, D. Hannequin, A. Laquerriere, V. Golfier, D. Seilhean, G. Viennet, P. Couratier, P. Verpillat, S. Heath, W. Camu, O. Martinaud, L. Lacomblez, M. Vercelletto, F. Salachas, F. Sellal, M. Didic, C. Thomas-Anterion, M. Puel, B. F. Michel, C. Besse, C. Duyckaerts, V. Meininger, D. Campion, B. Dubois, A. Brice, A. Brice, F. Blanc, W. Camu, F. Clerget-Darpoux, P. Corcia, M. Didic, V. de la Sayette, C. Desnuelle, B. Dubois, C. Duyckaerts, M. O. Habert, E. Guedj, D. Hannequin, L. Lacomblez, I. Le Ber, R. Levy, V. Meininger, B. F. Michel, F. Pasquier, C. Thomas-Anterion, M. Puel, F. Salachas, F. Sellal, M. Vercelletto, and P. Verpillat. Chromosome 9p-linked families with frontotemporal dementia associated with motor neuron disease. *Neurology*, 72(19):1669–1676, May 2009. 19

H. Lee, J. L. Deignan, N. Dorrani, S. P. Strom, S. Kantarci, F. Quintero-Rivera, K. Das, T. Toy, B. Harry, M. Yourshaw, M. Fox, B. L. Fogel, J. A. Martinez-Agosto, D. A. Wong, V. Y. Chang, P. B. Shieh, C. G. Palmer, K. M. Dipple, W. W. Grody, E. Vilain, and S. F. Nelson. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*, 312(18):1880–1887, Nov 2014. 32

J. A. Lee and J. R. Lupski. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, 52(1):103–121, Oct 2006. 283

Y. B. Lee, H. J. Chen, J. N. Peres, J. Gomez-Deza, J. Attig, M. Stalekar, C. Troakes, A. L. Nishimura, E. L. Scotter, C. Vance, Y. Adachi, V. Sardone, J. W. Miller, B. N. Smith, J. M. Gallo, J. Ule, F. Hirth, B. Rogelj, C. Houart, and C. E. Shaw.

Hexanucleotide repeats in ALS/FTD form length-dependent RNA foci, sequester RNA binding proteins, and are neurotoxic. *Cell Rep*, 5(5):1178–1186, Dec 2013. 22

Y. C. Lee, A. Durr, K. Majczenko, Y. H. Huang, Y. C. Liu, C. C. Lien, P. C. Tsai, Y. Ichikawa, J. Goto, M. L. Monin, J. Z. Li, M. Y. Chung, E. Mundwiller, V. Shakkottai, T. T. Liu, C. Tesson, Y. C. Lu, A. Brice, S. Tsuji, M. Burmeister, G. Stevanin, and B. W. Soong. Mutations in KCND3 cause spinocerebellar ataxia type 22. *Ann. Neurol.*, 72(6):859–869, Dec 2012. 32

P. N. Leigh, A. Dodson, M. Swash, J. P. Brion, and B. H. Anderton. Cytoskeletal abnormalities in motor neuron disease. An immunocytochemical study. *Brain*, 112 (Pt 2):521–535, Apr 1989. 7

M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson,

M. J. Daly, D. G. MacArthur, H. E. Abboud, G. Abecasis, C. A. Aguilar-Salinas, O. Arellano-Campos, G. Atzmon, I. Aukrust, C. L. Barr, G. I. Bell, G. I. Bell, S. Bergen, L. Bjorkhaug, J. Blangero, D. W. Bowden, C. L. Budman, N. P. Burtt, F. Centeno-Cruz, J. C. Chambers, K. Chambert, R. Clarke, R. Collins, G. Coppola, E. J. Cordova, M. L. Cortes, N. J. Cox, R. Duggirala, M. Farrall, J. C. Fernandez-Lopez, P. Fontanillas, T. M. Frayling, N. B. Freimer, C. Fuchsberger, H. Garcia-Ortiz, A. Goel, M. J. Gomez-Vazquez, M. E. Gonzalez-Villalpando, C. Gonzalez-Villalpando, M. A. Grados, L. Groop, C. A. Haiman, C. L. Hanis, C. L. Hanis, A. T. Hattersley, B. E. Henderson, J. C. Hopewell, A. Huerta-Chagoya, S. Islas-Andrade, S. B. Jacobs, S. Jalilzadeh, C. P. Jenkinson, J. Moran, S. Jimenez-Morale, A. Kahler, R. A. King, G. Kirov, J. S. Kooner, T. Kyriakou, J. Y. Lee, D. M. Lehman, G. Lyon, W. MacMahon, P. K. Magnusson, A. Mahajan, J. Marrugat, A. Martinez-Hernandez, C. A. Mathews, G. McVean, J. B. Meigs, T. Meitinger, E. Mendoza-Caamal, J. M. Mercader, K. L. Mohlke, H. Moreno-Macias, A. P. Morris, L. A. Najmi, P. R. Njolstad, M. C. O'Donovan, M. L. Ordonez-Sanchez, M. J. Owen, T. Park, D. L. Pauls, D. Posthuma, C. Revilla-Monsalve, L. Riba, S. Ripke, R. Rodriguez-Guillen, M. Rodriguez-Torres, P. Sandor, M. Seielstad, R. Sladek, X. Soberon, T. D. Spector, S. E. Tai, T. M. Teslovich, G. Walford, L. R. Wilkens, and A. L. Williams. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 08 2016. 31, 54, 56, 57, 184

S. H. Lelieveld, M. Spielmann, S. Mundlos, J. A. Veltman, and C. Gilissen. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum. Mutat.*, 36(8):815–822, Aug 2015. 286, 287

S. H. Lelieveld, J. A. Veltman, and C. Gilissen. Novel bioinformatic developments for exome sequencing. *Hum. Genet.*, 135(6):603–614, 06 2016. 31

I. Letunic, T. Doerks, and P. Bork. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, 43(Database issue):D257–260, Jan 2015. 56, 57

H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, Nov 2011. 64, 207

H. Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, Oct 2014. 268, 270, 291, 292, 293, 295

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009. 48, 225, 288

H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010. 48, 225, 288

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009. 48

J. Li, A. M. Batcha, B. Gruning, and U. R. Mansmann. An NGS Workflow Blueprint for DNA Sequencing Data and Its Application in Individualized Molecular Oncology. *Cancer Inform*, 14(Suppl 5):87–107, 2015. 288

X. L. Li, S. Shu, X. G. Li, Q. Liu, F. Liu, B. Cui, M. S. Liu, B. Peng, L. Y. Cui, and X. Zhang. CHCHD10 is not a frequent causative gene in Chinese ALS patients. *Amyotroph Lateral Scler Frontotemporal Degener*, 17(5-6):458–460, 2016. 106

M. D. Linderman, T. Brandt, L. Edelmann, O. Jabado, Y. Kasai, R. Kornreich, M. Mahajan, H. Shah, A. Kasarskis, and E. E. Schadt. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics*, 7:20, Apr 2014. 291, 293

S. C. Ling, M. Polymenidou, and D. W. Cleveland. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron*, 79(3):416–438, Aug 2013. 8

S. Liu, Y. Zhang, H. Bian, and X. Li. Gene expression profiling predicts pathways and genes associated with Parkinson's disease. *Neurol. Sci.*, 37(1):73–79, Jan 2016. 215

X. Liu, X. Jian, and E. Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, 32(8):894–899, Aug 2011. 49, 289

X. Liu, S. Han, Z. Wang, J. Gelernter, and B. Z. Yang. Variant callers for next-generation sequencing data: a comparison study. *PLoS ONE*, 8(9):e75619, 2013. 270, 288, 289, 290, 293, 295

Y. T. Liu, Y. C. Lee, and B. W. Soong. What we have learned from the next-generation sequencing: Contributions to the genetic diagnoses and understanding of pathomechanisms of neurodegenerative diseases. *J. Neurogenet.*, 29(2-3):103–112, 2015. 31, 32, 33, 286, 287

K. Lohmann and C. Klein. Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics*, 11(4):699–707, Oct 2014. 268, 289, 293, 294, 296

C. Lomen-Hoerth, T. Anderson, and B. Miller. The overlap of amyotrophic lateral sclerosis and frontotemporal dementia. *Neurology*, 59(7):1077–1079, Oct 2002. 17

A. A. Luty, J. B. Kwok, E. M. Thompson, P. Blumbergs, W. S. Brooks, C. T. Loy, C. Dobson-Stone, P. K. Panegyres, J. Hecker, G. A. Nicholson, G. M. Halliday, and P. R. Schofield. Pedigree with frontotemporal lobar degeneration–motor neuron disease and Tar DNA binding protein-43 positive neuropathology: genetic linkage to chromosome 9. *BMC Neurol*, 8:32, 2008. 19

D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476, Apr 2014. 33, 183, 211, 278, 296, 297, 298, 299, 300

I. R. Mackenzie, E. H. Bigio, P. G. Ince, F. Geser, M. Neumann, N. J. Cairns, L. K. Kwong, M. S. Forman, J. Ravits, H. Stewart, A. Eisen, L. McClusky, H. A. Kretzschmar, C. M. Monoranu, J. R. Highley, J. Kirby, T. Siddique, P. J. Shaw, V. M. Lee, and J. Q. Trojanowski. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Ann. Neurol.*, 61(5):427–434, May 2007. 15

I. R. Mackenzie, T. Arzberger, E. Kremmer, D. Troost, S. Lorenzl, K. Mori, S. M. Weng, C. Haass, H. A. Kretzschmar, D. Edbauer, and M. Neumann. Dipeptide repeat protein pathology in C9ORF72 mutation cases: clinico-pathological correlations. *Acta Neuropathol.*, 126(6):859–879, Dec 2013. 22

I. R. Mackenzie, A. M. Nicholson, M. Sarkar, J. Messing, M. D. Purice, C. Pottier, K. Annu, M. Baker, R. B. Perkerson, A. Kurti, B. J. Matchett, T. Mittag, J. Temirov, G. R. Hsiung, C. Krieger, M. E. Murray, M. Kato, J. D. Fryer, L. Petrucelli, L. Zinman, S. Weintraub, M. Mesulam, J. Keith, S. A. Zivkovic, V. Hirsch-Reinshagen, R. P. Roos, S. Zuchner, N. R. Graff-Radford, R. C. Petersen, R. J. Caselli, Z. K. Wszolek, E. Finger, C. Lippa, D. Lacomis, H. Stewart, D. W. Dickson, H. J. Kim, E. Rogaeva, E. Bigio, K. B. Boylan, J. P. Taylor, and R. Rademakers. TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron*, 95(4):808–816, Aug 2017. 12, 25, 35, 154, 160, 161

L. MacNair, S. Xiao, D. Miletic, M. Ghani, J. P. Julien, J. Keith, L. Zinman, E. Rogaeva, and J. Robertson. MTHFSD and DDX58 are novel RNA-binding proteins abnormally regulated in amyotrophic lateral sclerosis. *Brain*, 139(Pt 1):86–100, Jan 2016. 154

A. Mahto. *splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values*, 2017. URL `https://CRAN.R-project.org/package=readr`. 65

E. Majounie, A. E. Renton, K. Mok, E. Dopper, A. Waite, S. Rollinson, A. Chi, G. Restagno, N. Nicolaou, J. Simon-Sanchez, J. C. van Swieten, Y. Abramzon, J. O. Johnson, M. Sendtner, R. Pamphlett, R. W. Orrell, S. Mead, K. Sidle, H. Houlden, J. Rohrer, K. E. Morrison, H. Pall, K. Talbot, O. Ansorge, D. Hernandez, S. Arepalli, M. Sabatelli, G. Mora, M. Corbo, F. Giannini, A. Calvo, E. Englund, G. Borghero, G. L. Floris, A. Remes, H. Laaksovirta, L. McCluskey, J. Q. Trojanowski, V. M. Van Deerlin, G. D. Schellenberg, M. A. Nalls, V. E. Drory, C. S. Lu, T. H. Yeh, H. Ishiura, Y. Takahashi, S. Tsuji, I. Le Ber, A. Brice, C. Drepper, N. Williams, J. Kirby, P. Shaw, J. Hardy, P. J. Tienari, P. Heutink, H. R. Morris, S. Pickering-Brown, and B. J. Traynor. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurol*, 11(4):323–330, Apr 2012. 20, 21

A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, Nov 2010. 181

M. A. Mansilla, J. Kimani, L. E. Mitchell, K. Christensen, D. I. Boomsma, S. Daack-Hirsch, B. Nepomucena, D. F. Wyszynski, T. M. Felix, N. G. Martin, and J. C. Murray. Discordant MZ twins with cleft lip and palate: a model for identifying genes in complex traits. *Twin Res Hum Genet*, 8(1):39–46, Feb 2005. 267, 279

B. Marin, F. Boumediene, G. Logroscino, P. Couratier, M. C. Babron, A. L. Leutenegger, M. Copetti, P. M. Preux, and E. Beghi. Variation in worldwide incidence of amyotrophic lateral sclerosis: a meta-analysis. *Int J Epidemiol*, 46(1):57–74, 02 2017. 5

N. Marroquin, S. Stranz, K. Muller, T. Wieland, W. P. Ruf, S. J. Brockmann, K. M. Danzer, G. Borck, A. Hubers, P. Weydt, T. Meitinger, T. M. Strom, A. Rosenbohm, A. C. Ludolph, and J. H. Weishaupt. Screening for CHCHD10 mutations in a large cohort of sporadic ALS patients: no evidence for pathogenicity of the p.P34S variant. *Brain*, 139(Pt 2):e8, Feb 2016. 106

H. Maruyama, H. Morino, H. Ito, Y. Izumi, H. Kato, Y. Watanabe, Y. Kinoshita, M. Kamada, H. Nodera, H. Suzuki, O. Komure, S. Matsuura, K. Kobatake, N. Morimoto, K. Abe, N. Suzuki, M. Aoki, A. Kawata, T. Hirai, T. Kato, K. Ogasawara, A. Hirano, T. Takumi, H. Kusaka, K. Hagiwara, R. Kaji, and H. Kawakami. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature*, 465(7295):223–226, May 2010. 10, 12, 25

I. Mathieson and G. McVean. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, 44(3):243–246, Feb 2012. 301

H. Matsumine, M. Saito, S. Shimoda-Matsubayashi, H. Tanaka, A. Ishikawa, Y. Nakagawa-Hattori, M. Yokochi, T. Kobayashi, S. Igarashi, H. Takano, K. Sanpei, R. Koike, H. Mori, T. Kondo, Y. Mizutani, A. A. Schaffer, Y. Yamamura, S. Nakamura, S. Kuzuhara, S. Tsuji, and Y. Mizuno. Localization of a gene for an autosomal recessive form of juvenile Parkinsonism to chromosome 6q25.2-27. *Am. J. Hum. Genet.*, 60(3):588–596, Mar 1997. 154

E. P. McCann, K. L. Williams, J. A. Fifita, I. S. Tarr, J. O'Connor, D. B. Rowe, G. A. Nicholson, and I. P. Blair. The genotype-phenotype landscape of familial amyotrophic lateral sclerosis in Australia. *Clin. Genet.*, 92(3):259–266, Sep 2017. 11, 13, 14, 18, 21, 168

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, Sep 2010. 48, 225, 288

R. L. McLaughlin, K. P. Kenna, A. Vajda, S. Byrne, D. G. Bradley, and O. Hardiman. UBQLN2 mutations are not a frequent cause of amyotrophic lateral sclerosis in Ireland. *Neurobiol. Aging*, 35(1):9–11, Jan 2014. 23

R. L. McLaughlin, A. Vajda, and O. Hardiman. Heritability of Amyotrophic Lateral Sclerosis: Insights From Disparate Numbers. *JAMA Neurol*, 72(8):857–858, Aug 2015. 26, 223

J. Meienberg, R. Bruggmann, K. Oexle, and G. Matyas. Clinical sequencing: is WGS the better WES? *Hum. Genet.*, 135(3):359–362, Mar 2016. 268, 270, 287

K. Meltz Steinberg, T. J. Nicholas, D. C. Koboldt, B. Yu, E. Mardis, and R. Pamphlett. Whole genome analyses reveal no pathogenetic single nucleotide or structural

differences between monozygotic twins discordant for amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener*, 16(5-6):385–392, 2015. 267, 279

M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11 (1):31–46, Jan 2010. 45

A. M. Meynert, L. S. Bicknell, M. E. Hurles, A. P. Jackson, and M. S. Taylor. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics*, 14:195, Jun 2013. 287

A. M. Meynert, M. Ansari, D. R. FitzPatrick, and M. S. Taylor. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, 15:247, Jul 2014. 286, 287, 292

A. M. Middleton. Hpcc systems: Introduction to hpcc (high-performance computing cluster), May 2011. 62

M. Mielczarek and J. Szyda. Review of alignment and SNP calling algorithms for next-generation sequencing data. *J. Appl. Genet.*, 57(1):71–79, Feb 2016. 288, 289, 293

S. Millecamps, F. Salachas, C. Cazeneuve, P. Gordon, B. Bricka, A. Camuzat, L. Guillot-Noel, O. Russaouen, G. Bruneteau, P. F. Pradat, N. Le Forestier, N. Vandenberghe, V. Danel-Brunaud, N. Guy, C. Thauvin-Robinet, L. Lacomblez, P. Couratier, D. Hannequin, D. Seilhean, I. Le Ber, P. Corcia, W. Camu, A. Brice, G. Rouleau, E. LeGuern, and V. Meininger. SOD1, ANG, VAPB, TARDBP, and FUS mutations in familial amyotrophic lateral sclerosis: genotype-phenotype correlations. *J. Med. Genet.*, 47(8):554–560, Aug 2010. 18

S. Millecamps, P. Corcia, C. Cazeneuve, S. Boillee, D. Seilhean, V. Danel-Brunaud, N. Vandenberghe, P. F. Pradat, N. Le Forestier, L. Lacomblez, G. Bruneteau, W. Camu, A. Brice, V. Meininger, E. LeGuern, and F. Salachas. Mutations in UBQLN2 are rare in French amyotrophic lateral sclerosis. *Neurobiol. Aging*, 33(4): 1–3, Apr 2012. 23

R. G. Miller, J. D. Mitchell, M. Lyon, and D. H. Moore. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database Syst Rev*, (1):CD001447, Jan 2007. 6

T. M. Miller, A. Pestronk, W. David, J. Rothstein, E. Simpson, S. H. Appel, P. L. Andres, K. Mahoney, P. Allred, K. Alexander, L. W. Ostrow, D. Schoenfeld, E. A.

Macklin, D. A. Norris, G. Manousakis, M. Crisp, R. Smith, C. F. Bennett, K. M. Bishop, and M. E. Cudkowicz. An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial amyotrophic lateral sclerosis: a phase 1, randomised, first-in-man study. *Lancet Neurol*, 12(5):435–442, May 2013. 7

J. Mitchell, P. Paul, H. J. Chen, A. Morris, M. Payling, M. Falchi, J. Habgood, S. Panoutsou, S. Winkler, V. Tisato, A. Hajitou, B. Smith, C. Vance, C. Shaw, N. D. Mazarakis, and J. de Belleroche. Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proc. Natl. Acad. Sci. U.S.A.*, 107(16):7556–7561, Apr 2010. 12

S. Mizielinska and A. M. Isaacs. C9orf72 amyotrophic lateral sclerosis and frontotemporal dementia: gain or loss of function? *Curr. Opin. Neurol.*, 27(5):515–523, Oct 2014. 22

S. Mizielinska, S. Gronke, T. Niccoli, C. E. Ridler, E. L. Clayton, A. Devoy, T. Moens, F. E. Norona, I. O. C. Woollacott, J. Pietrzyk, K. Cleverley, A. J. Nicoll, S. Pickering-Brown, J. Dols, M. Cabecinha, O. Hendrich, P. Fratta, E. M. C. Fisher, L. Partridge, and A. M. Isaacs. C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. *Science*, 345(6201):1192–1194, Sep 2014. 21, 22

M. Mohiyuddin, J. C. Mu, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. Lam. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, 31(16):2741–2744, Aug 2015. 305

K. Mok, B. J. Traynor, J. Schymick, P. J. Tienari, H. Laaksovirta, T. Peuralinna, L. Myllykangas, A. Chio, A. Shatunov, B. F. Boeve, A. L. Boxer, M. DeJesus-Hernandez, I. R. Mackenzie, A. Waite, N. Williams, H. R. Morris, J. Simon-Sanchez, J. C. van Swieten, P. Heutink, G. Restagno, G. Mora, K. E. Morrison, P. J. Shaw, P. S. Rollinson, A. Al-Chalabi, R. Rademakers, S. Pickering-Brown, R. W. Orrell, M. A. Nalls, and J. Hardy. Chromosome 9 ALS and FTD locus is probably derived from a single founder. *Neurobiol. Aging*, 33(1):3–8, Jan 2012. 19

R. S. Moller, H. A. Dahl, and I. Helbig. The contribution of next generation sequencing to epilepsy genetics. *Expert Rev. Mol. Diagn.*, 15(12):1531–1538, 2015. 30, 31

Z. Monahan, F. Shewmaker, and U. B. Pandey. Stress granules at the intersection of autophagy and ALS. *Brain Res.*, 1649(Pt B):189–200, Oct 2016. 9

A. Montuschi, B. Iazzolino, A. Calvo, C. Moglia, L. Lopiano, G. Restagno, M. Brunetti, I. Ossola, A. Lo Presti, S. Cammarosano, A. Canosa, and A. Chio. Cognitive correlates in amyotrophic lateral sclerosis: a population-based study in Italy. *J. Neurol. Neurosurg. Psychiatry*, 86(2):168–173, Feb 2015. 5

J. M. Morahan, B. Yu, R. J. Trent, and R. Pamphlett. A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 10(5-6):418–429, 2009. 284

K. Mori, T. Arzberger, F. A. Grasser, I. Gijselinck, S. May, K. Rentzsch, S. M. Weng, M. H. Schludi, J. van der Zee, M. Cruts, C. Van Broeckhoven, E. Kremmer, H. A. Kretzschmar, C. Haass, and D. Edbauer. Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. *Acta Neuropathol.*, 126(6):881–893, Dec 2013a. 9, 22

K. Mori, S. M. Weng, T. Arzberger, S. May, K. Rentzsch, E. Kremmer, B. Schmid, H. A. Kretzschmar, M. Cruts, C. Van Broeckhoven, C. Haass, and D. Edbauer. The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. *Science*, 339(6125):1335–1338, Mar 2013b. 9, 22

M. Morita, A. Al-Chalabi, P. M. Andersen, B. Hosler, P. Sapp, E. Englund, J. E. Mitchell, J. J. Habgood, J. de Belleroche, J. Xi, W. Jongjaroenprasert, H. R. Horvitz, L. G. Gunnarsson, and R. H. Brown. A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology*, 66(6):839–844, Mar 2006. 19

N. E. Morton. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*, 7(3): 277–318, Sep 1955. 30

S. I. Mota, I. L. Ferreira, and A. C. Rego. Dysfunctional synapse in Alzheimer's disease - A focus on NMDA receptors. *Neuropharmacology*, 76 Pt A:16–26, Jan 2014. 215

J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, 19(R2):R131–136, Oct 2010. 226, 270

K. Muller, P. M. Andersen, A. Hubers, N. Marroquin, A. E. Volk, K. M. Danzer, T. Meitinger, A. C. Ludolph, T. M. Strom, and J. H. Weishaupt. Two novel mutations in conserved codons indicate that CHCHD10 is a gene associated with motor neuron disease. *Brain*, 137(Pt 12):e309, Dec 2014. 106, 143

S. Nakanishi. Molecular diversity of glutamate receptors and implications for brain function. *Science*, 258(5082):597–603, Oct 1992. 215

M. Neumann, D. M. Sampathu, L. K. Kwong, A. C. Truax, M. C. Micsenyi, T. T. Chou, J. Bruce, T. Schuck, M. Grossman, C. M. Clark, L. F. McCluskey, B. L. Miller, E. Masliah, I. R. Mackenzie, H. Feldman, W. Feiden, H. A. Kretzschmar, J. Q. Trojanowski, and V. M. Lee. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, 314(5796):130–133, Oct 2006. 8, 11, 15, 168

M. Neumann, E. Bentmann, D. Dormann, A. Jawaid, M. DeJesus-Hernandez, O. Ansorge, S. Roeber, H. A. Kretzschmar, D. G. Munoz, H. Kusaka, O. Yokota, L. C. Ang, J. Bilbao, R. Rademakers, C. Haass, and I. R. Mackenzie. FET proteins TAF15 and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. *Brain*, 134(Pt 9):2595–2609, Sep 2011. 8, 168

K. Neveling, I. Feenstra, C. Gilissen, L. H. Hoefsloot, E. J. Kamsteeg, A. R. Mensenkamp, R. J. Rodenburg, H. G. Yntema, L. Spruijt, S. Vermeer, T. Rinne, K. L. van Gassen, D. Bodmer, D. Lugtenberg, R. de Reuver, W. Buijsman, R. C. Derks, N. Wieskamp, B. van den Heuvel, M. J. Ligtenberg, H. Kremer, D. A. Koolen, B. P. van de Warrenburg, F. P. Cremers, C. L. Marcelis, J. A. Smeitink, S. B. Wortmann, W. A. van Zelst-Stams, J. A. Veltman, H. G. Brunner, H. Scheffer, and M. R. Nelen. A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum. Mutat.*, 34(12):1721–1726, Dec 2013a. 32

K. Neveling, L. A. Martinez-Carrera, I. Holker, A. Heister, A. Verrips, S. M. Hosseini-Barkooie, C. Gilissen, S. Vermeer, M. Pennings, R. Meijer, M. te Riele, C. J. Frijns, O. Suchowersky, L. MacLaren, S. Rudnik-Schoneborn, R. J. Sinke, K. Zerres, R. B. Lowry, H. H. Lemmink, L. Garbes, J. A. Veltman, H. J. Schelhaas, H. Scheffer, and B. Wirth. Mutations in BICD2, which encodes a golgin and important motor adaptor, cause congenital autosomal-dominant spinal muscular atrophy. *Am. J. Hum. Genet.*, 92(6):946–954, Jun 2013b. 154

A. S. L. Ng and E. K. Tan. Intermediate C9orf72 alleles in neurological disorders: does size really matter? *J. Med. Genet.*, 54(9):591–597, Sep 2017. 20, 21

S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, Sep 2009. 31

S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, and J. Shendure. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, 42(9):790–793, Sep 2010a. 32

S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, 42(1): 30–35, Jan 2010b. 32

D. K. H. Nguyen, R. Thombre, and J. Wang. Autophagy as a common pathway in amyotrophic lateral sclerosis. *Neurosci. Lett.*, Apr 2018a. 24

H. P. Nguyen, C. Van Broeckhoven, and J. van der Zee. ALS Genes in the Genomic Era and their Implications for FTD. *Trends Genet.*, 34(6):404–423, Jun 2018b. 280

H. P. Nguyen, S. Van Mossevelde, L. Dillen, J. L. De Bleecker, M. Moisse, P. Van Damme, C. Van Broeckhoven, J. van der Zee, S. Engelborghs, R. Crols, P. P. De Deyn, P. De Jonghe, J. Baets, P. Cras, R. Mercelis, R. Vandenberghe, A. Sieben, P. Santens, A. Ivanoiu, O. Deryck, L. Vanopdenbosch, and J. Delbeck. NEK1 genetic variability in a Belgian cohort of ALS and ALS-FTD patients. *Neurobiol. Aging*, 61: 1–255, 01 2018c. 280

A. Nicolas. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron*, 97 (6):1268–1283, 2018. 154

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, 12(6):443–451, Jun 2011. 292, 293

A. Niroula and M. Vihinen. Harmful somatic amino acid substitutions affect key pathways in cancers. *BMC Med Genomics*, 8:53, 2015. 57

A. L. Nishimura, M. Mitne-Neto, H. C. Silva, A. Richieri-Costa, S. Middleton, D. Cascio, F. Kok, J. R. Oliveira, T. Gillingwater, J. Webb, P. Skehel, and M. Zatz. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *Am. J. Hum. Genet.*, 75(5):822–831, Nov 2004. 12, 25

A. Nishiyama, T. Niihori, H. Warita, R. Izumi, T. Akiyama, M. Kato, N. Suzuki, Y. Aoki, and M. Aoki. Comprehensive targeted next-generation sequencing in Japanese familial amyotrophic lateral sclerosis. *Neurobiol. Aging*, 53:1–194, 05 2017. 13, 20

T. Nonaka, F. Kametani, T. Arai, H. Akiyama, and M. Hasegawa. Truncation and pathogenic mutations facilitate the formation of intracellular aggregates of TDP-43. *Hum. Mol. Genet.*, 18(18):3353–3364, Sep 2009. 16

E. C. Oates, A. M. Rossor, M. Hafezparast, M. Gonzalez, F. Speziani, D. G. MacArthur, M. Lek, E. Cottenie, M. Scoto, A. R. Foley, M. Hurles, H. Houlden, L. Greensmith, M. Auer-Grumbach, T. R. Pieber, T. M. Strom, R. Schule, D. N. Herrmann, J. E. Sowden, G. Acsadi, M. P. Menezes, N. F. Clarke, S. Zuchner, F. Muntoni, K. N. North, and M. M. Reilly. Mutations in BICD2 cause dominant congenital spinal muscular atrophy and hereditary spastic paraplegia. *Am. J. Hum. Genet.*, 92(6): 965–973, Jun 2013. 154

V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, and M. Morgan. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14):2076–2078, Jul 2014. 65

T. D. O'Connor, A. Kiezun, M. Bamshad, S. S. Rich, J. D. Smith, E. Turner, S. M. Leal, J. M. Akey, S. B. Gabriel, D. M. Altshuler, G. R. Abecasis, H. Allayee, S. Cresci, M. J. Daly, P. I. de Bakker, M. A. Depristo, R. Do, P. Donnelly, D. N. Farlow, T. Fennell, K. Garimella, S. L. Hazen, Y. Hu, D. M. Jordan, G. Jun, S. Kathiresan, S. Kawut, A. Kiezun, G. Kryukov, G. Lettre, B. Li, M. Li, C. H. Newton-Cheh, S. Padmanabhan, S. Pulit, D. J. Rader, D. Reich, M. P. Reilly, M. A. Rivas, S. Schwartz, L. Scott, J. A. Spertus, N. O. Stitziel, N. Stoletzki, S. R. Sunyaev, B. F. Voight, C. J. Willer, S. S. Rich, E. Akylbekova, L. D. Atwood, C. M. Ballantyne, M. Barbalic, R. Barr, E. J. Benjamin, J. Bis, C. Bizon, E. Boerwinkle, D. W. Bowden, J. Brody, M. Budoff, G. Burke, S. Buxbaum, J. Carr, D. T. Chen, I. Y. Chen, W. M. Chen, P. Concannon, J. Crosby, L. Cupples, R. D'Agostino, A. L. DeStefano, A. Dreisbach, J. Dupuis, J. Durda, J. Ellis, A. R. Folsom, M. Fornage, C. S. Fox, E. Fox, V. Funari, S. K. Ganesh, J. Gardin, D. Goff, O. Gordon, W. Grody, M. Gross, X. Guo, I. M. Hall, N. L. Heard-Costa, S. R. Heckbert, N. Heintz, D. M. Herrington, D. Hickson, J. Huang, S. J. Hwang, D. R. Jacobs, N. S. Jenny, A. D. Johnson, C. W. Johnson, R. Kronmal, R. Kurz, E. M. Lange, L. A. Lange, M. G. Larson, M. Lawson, D. Levy, D. Li, H. Lin, C. Liu, J. Liu, K. Liu, X. Liu, Y. Liu, W. T. Longstreth, C. Loria, T. Lumley, K. Lunetta, A. J. Mackey, R. Mackey,

A. Manichaikul, T. Maxwell, B. McKnight, J. B. Meigs, A. C. Morrison, S. K. Musani, J. C. Mychaleckyj, J. A. Nettleton, K. North, C. J. O'Donnell, D. O'Leary, F. Ong, W. Palmas, J. S. Pankow, N. D. Pankratz, S. Paul, M. Perez, S. D. Person, J. Polak, W. S. Post, B. M. Psaty, A. R. Quinlan, L. J. Raffel, V. S. Ramachandran, A. P. Reiner, K. Rice, J. I. Rotter, J. P. Sanders, P. Schreiner, S. Seshadri, S. Shea, S. Sidney, K. Silverstein, D. S. Siscovick, N. L. Smith, N. Sotoodehnia, A. Srinivasan, H. A. Taylor, K. Taylor, F. Thomas, R. P. Tracy, M. Y. Tsai, K. A. Volcik, C. L. Wassel, K. Watson, G. Wei, W. White, K. L. Wiggins, J. B. Wilk, O. Williams, J. G. Wilson, P. Wolf, N. A. Zakai, J. Hardy, J. F. Meschia, M. Nalls, S. S. Rich, A. Singleton, B. Worrall, M. J. Bamshad, K. C. Barnes, I. Abdulhamid, F. Accurso, R. Anbar, T. Beaty, A. Bigham, P. Black, E. Bleecker, K. Buckingham, A. M. Cairns, D. Caplan, B. Chatfield, A. Chidekel, M. Cho, D. C. Christiani, J. D. Crapo, J. Crouch, D. Daley, A. Dang, H. Dang, A. De Paula, J. DeCelie-Germana, A. Dozor, M. Drumm, M. Dyson, J. Emerson, M. J. Emond, T. Ferkol, R. Fink, C. Foster, D. Froh, L. Gao, W. Gershan, R. L. Gibson, E. Godwin, M. Gondor, H. Gutierrez, N. N. Hansel, P. M. Hassoun, P. Hiatt, J. E. Hokanson, M. Howenstine, L. K. Hummer, J. Kanga, Y. Kim, M. R. Knowles, M. Konstan, T. Lahiri, N. Laird, C. Lange, L. Lin, T. L. Louie, D. Lynch, B. Make, T. R. Martin, S. C. Mathai, R. A. Mathias, J. McNamara, S. McNamara, D. Meyers, S. Millard, P. Mogayzel, R. Moss, T. Murray, D. Nielson, B. Noyes, W. O'Neal, D. Orenstein, B. O'Sullivan, R. Pace, P. Pare, H. Parker, M. A. Passero, E. Perkett, A. Prestridge, N. M. Rafaels, B. Ramsey, E. Regan, C. Ren, G. Retsch-Bogart, M. Rock, A. Rosen, M. Rosenfeld, I. Ruczinski, A. Sanford, D. Schaeffer, C. Sell, D. Sheehan, E. K. Silverman, D. Sin, T. Spencer, J. Stonebraker, H. K. Tabor, L. Varlotta, C. I. Vergara, R. Weiss, F. Wigley, R. A. Wise, F. A. Wright, M. M. Wurfel, R. Zanni, F. Zou, D. A. Nickerson, M. J. Rieder, P. Green, J. Shendure, B. Paeper, J. M. Akey, M. J. Bamshad, C. D. Bustamante, D. R. Crosslin, E. E. Eichler, P. Fox, A. Gordon, S. Gravel, G. P. Jarvik, J. M. Johnsen, E. E. Kenny, J. M. Kidd, F. Lara-Garduno, S. M. Leal, D. J. Liu, S. McGee, P. D. Robertson, J. D. Smith, J. C. Staples, E. H. Turner, G. Wang, R. Jackson, K. North, U. Peters, C. S. Carlson, G. Anderson, H. Anton-Culver, T. L. Assimes, P. L. Auer, S. Beresford, C. Bizon, H. Black, R. Brunner, R. Brzyski, D. Burwen, B. Caan, C. L. Carty, R. Chlebowski, S. Cummings, J. Curb, C. B. Eaton, L. Ford, N. Franceschini, S. M. Fullerton, M. Gass, N. Geller, G. Heiss, B. V. Howard, L. Hsu, C. M. Hutter, J. Ioannidis, S. Jiao, K. C. Johnson, E. Kabagambe, C. Kooperberg, L. Kuller, A. LaCroix, K. Lakshminarayan, D. Lane, E. M. Lange, L. A. Lange, N. Lasser, E. LeBlanc, C. E. Lewis, M. Limacher, D. Lin, B. A. Logsdon, S. Ludlam, J. E. Manson, K. Margolis, L. Martin, J. McGowan, K. L. Monda, J. M. Kotchen,

L. Nathan, J. Ockene, M. J. O'Sullivan, L. S. Phillips, R. L. Prentice, A. P. Reiner, J. Robbins, J. G. Robinson, J. E. Rossouw, H. Sangi-Haghpeykar, G. E. Sarto, S. Shumaker, M. S. Simon, M. L. Stefanick, E. Stein, H. Tang, K. C. Taylor, C. A. Thomson, T. A. Thornton, L. Van Horn, M. Vitolins, J. Wactawski-Wende, R. Wallace, S. Wassertheil-Smoller, D. Applebaum-Bowden, M. Feolo, W. Gan, W. Hoots, J. Kiley, M. Lauer, H. Leeds, A. Michelson, D. N. Paltoo, P. Sholinsky, S. Skarlatos, and A. Sturcke. Fine-scale patterns of population stratification confound rare variant association tests. *PLoS ONE*, 8(7):e65834, 2013. 301

S. O'Dowd, D. Curtin, A. J. Waite, K. Roberts, N. Pender, V. Reid, M. O'Connell, N. M. Williams, H. R. Morris, B. J. Traynor, and T. Lynch. C9ORF72 expansion in amyotrophic lateral sclerosis/frontotemporal dementia also causes parkinsonism. *Mov. Disord.*, 27(8):1072–1074, Jul 2012. 20

Y. Ohki, A. Wenninger-Weinzierl, A. Hruscha, K. Asakawa, K. Kawakami, C. Haass, D. Edbauer, and B. Schmid. Glycine-alanine dipeptide repeat protein contributes to toxicity in a zebrafish model of C9orf72 associated neurodegeneration. *Mol Neurodegener*, 12(1):6, 01 2017. 22

M. O'Huallachain, K. J. Karczewski, S. M. Weissman, A. E. Urban, and M. P. Snyder. Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44):18018–18023, Oct 2012. 267

K. Okamoto, Y. Mizuno, and Y. Fujita. Bunina bodies in amyotrophic lateral sclerosis. *Neuropathology*, 28(2):109–115, Apr 2008. 8

J. W. Olney. Excitotoxic amino acids and neuropsychiatric disorders. *Annu. Rev. Pharmacol. Toxicol.*, 30:47–71, 1990. 215

J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*, 5(3):28, 2013. 270, 289, 293, 295

H. T. Orr, M. Y. Chung, S. Banfi, T. J. Kwiatkowski, A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. Ranum, and H. Y. Zoghbi. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, 4 (3):221–226, Jul 1993. 283, 305

M. Osaka, D. Ito, T. Yagi, Y. Nihei, and N. Suzuki. Evidence of a link between ubiquilin 2 and optineurin in amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 24 (6):1617–1629, Mar 2015. 24

M. Osaka, D. Ito, and N. Suzuki. Disturbance of proteasomal and autophagic protein degradation pathways by amyotrophic lateral sclerosis-linked mutations in ubiquilin 2. *Biochem. Biophys. Res. Commun.*, 472(2):324–331, Apr 2016. 24

J. Oshima, D. B. Magner, J. A. Lee, A. M. Breman, E. S. Schmitt, L. D. White, C. A. Crowe, M. Merrill, P. Jayakar, A. Rajadhyaksha, C. M. Eng, and D. del Gaudio. Regional genomic instability predisposes to complex dystrophin gene rearrangements. *Hum. Genet.*, 126(3):411–423, Sep 2009. 283, 305

B. Oskarsson, D. K. Horton, and H. Mitsumoto. Potential Environmental Factors in Amyotrophic Lateral Sclerosis. *Neurol Clin*, 33(4):877–888, Nov 2015. 6

J. Ott, J. Wang, and S. M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.*, 16(5):275–284, May 2015. 28, 29, 38

A. Ozoguz, O. Uyan, G. Birdal, C. Iskender, E. Kartal, S. Lahut, O. Omur, Z. S. Agim, A. G. Eken, N. E. Sen, P. Kavak, C. Sayg, P. C. Sapp, P. Keagle, Y. Parman, E. Tan, F. Koc, F. Deymeer, P. Oflazer, H. Hanagas, H. Gurvit, B. Bilgic, H. Durmus, M. Ertas, D. Kotan, M. A. Akaln, H. Glloglu, M. Zarifoglu, F. Aysal, N. Dsoglu, K. Bilguvar, M. Gunel, O. Keskin, T. Akgun, H. Ozcelik, J. E. Landers, R. H. Brown, and A. N. Basak. The distinct genetic pattern of ALS in Turkey and novel mutations. *Neurobiol. Aging*, 36(4):9–1764, Apr 2015. 23

S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M. R. Speicher, J. Zschocke, and Z. Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinformatics*, 15(2):256–278, Mar 2014. 52, 270, 271, 288, 295, 298

E. E. Palmer, T. Stuhlmann, S. Weinert, E. Haan, H. Van Esch, M. Holvoet, J. Boyle, M. Leffler, M. Raynaud, C. Moraine, H. van Bokhoven, T. Kleefstra, K. Kahrizi, H. Najmabadi, H. H. Ropers, M. R. Delgado, D. Sirsi, S. Golla, A. Sommer, M. P. Pietryga, W. K. Chung, J. Wynn, L. Rohena, E. Bernardo, D. Hamlin, B. M. Faux, D. K. Grange, L. Manwaring, J. Tolmie, S. Joss, J. M. Cobben, F. A. M. Duijkers, J. M. Goehringer, T. D. Challman, F. Hennig, U. Fischer, A. Grimme, V. Suckow, L. Musante, J. Nicholl, M. Shaw, S. P. Lodh, Z. Niu, J. A. Rosenfeld, P. Stankiewicz, T. J. Jentsch, J. Gecz, M. Field, and V. M. Kalscheuer. De novo and inherited

mutations in the X-linked gene CLCN4 are associated with syndromic intellectual disability and behavior and seizure disorders in males and females. *Mol. Psychiatry*, 23(2):222–230, Feb 2018. 214

N. Parkinson, P. G. Ince, M. O. Smith, R. Highley, G. Skibinski, P. M. Andersen, K. E. Morrison, H. S. Pall, O. Hardiman, J. Collinge, P. J. Shaw, and E. M. Fisher. ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B). *Neurology*, 67(6):1074–1077, Sep 2006. 10, 12

N. Pasquarelli, M. Engelskirchen, J. Hanselmann, S. Endres, C. Porazik, H. Bayer, E. Buck, M. Karsak, P. Weydt, B. Ferger, and A. Witting. Evaluation of monoacylglycerol lipase as a therapeutic target in a transgenic mouse model of ALS. *Neuropharmacology*, 124:157–169, Sep 2017. 154

J. P. Pearson, N. M. Williams, E. Majounie, A. Waite, J. Stott, V. Newsway, A. Murray, D. Hernandez, R. Guerreiro, A. B. Singleton, J. Neal, and H. R. Morris. Familial frontotemporal dementia with amyotrophic lateral sclerosis and a shared haplotype on chromosome 9p. *J. Neurol.*, 258(4):647–655, Apr 2011. 19

K. Peeters, I. Litvinenko, B. Asselbergh, L. Almeida-Souza, T. Chamova, T. Geuens, E. Ydens, M. Zimon, J. Irobi, E. De Vriendt, V. De Winter, T. Ooms, V. Timmerman, I. Tournev, and A. Jordanova. Molecular defects in the motor adaptor BICD2 cause proximal spinal muscular atrophy with autosomal-dominant inheritance. *Am. J. Hum. Genet.*, 92(6):955–964, Jun 2013. 154

F. Perrone, H. P. Nguyen, S. Van Mossevelde, M. Moisse, A. Sieben, P. Santens, J. De Bleecker, M. Vandenbulcke, S. Engelborghs, J. Baets, P. Cras, R. Vandenberghe, P. De Jonghe, P. P. De Deyn, J. J. Martin, P. Van Damme, C. Van Broeckhoven, J. van der Zee, D. Nuytten, K. Smets, J. Versijpt, A. Michotte, A. Ivanoiu, O. Deryck, B. Bergmans, J. Delbeck, M. Bruyland, C. Willems, and E. Salmon. Investigating the role of ALS genes CHCHD10 and TUBA4A in Belgian FTD-ALS spectrum patients. *Neurobiol. Aging*, 51:9–177, Mar 2017. 106, 143

B. S. Petersen, M. E. Spehlmann, A. Raedler, B. Stade, I. Thomsen, R. Rabionet, P. Rosenstiel, S. Schreiber, and A. Franke. Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. *BMC Genomics*, 15:564, 2014. 267, 279

D. Petrov, C. Mansfield, A. Moussy, and O. Hermine. ALS Clinical Trials Review: 20 Years of Failure. Are We Any Closer to Registering a New Treatment? *Front Aging Neurosci*, 9:68, 2017. 6

S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.*, 9(8):e1003709, 2013. 56, 57, 183

J. Phukan, M. Elamin, P. Bede, N. Jordan, L. Gallagher, S. Byrne, C. Lynch, N. Pender, and O. Hardiman. The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *J. Neurol. Neurosurg. Psychiatry*, 83(1):102–108, Jan 2012. 5

V. Picher-Martel, P. N. Valdmanis, P. V. Gould, J. P. Julien, and N. Dupre. From animal models to human disease: a genetic approach for personalized medicine in ALS. *Acta Neuropathol Commun*, 4(1):70, 07 2016. 303

M. Pirooznia, M. Kramer, J. Parla, F. S. Goes, J. B. Potash, W. R. McCombie, and P. P. Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics*, 8:14, Jul 2014. 288

M. Pletikos, A. M. Sousa, G. Sedmak, K. A. Meyer, Y. Zhu, F. Cheng, M. Li, Y. I. Kawasawa, and N. Sestan. Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron*, 81(2):321–332, Jan 2014. 55, 57, 195

K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, Jan 2010. 55, 57, 212

N. Popitsch, A. Schuh, and J. C. Taylor. ReliableGenome: annotation of genomic regions with high/low variant calling concordance. *Bioinformatics*, 33(2):155–160, 01 2017. 293, 295

D. S. Protter and R. Parker. Principles and Properties of Stress Granules. *Trends Cell Biol.*, 26(9):668–679, 09 2016. 9

I. Puls, C. Jonnakuty, B. H. LaMonte, E. L. Holzbaur, M. Tokito, E. Mann, M. K. Floeter, K. Bidus, D. Drayna, S. J. Oh, R. H. Brown, C. L. Ludlow, and K. H. Fischbeck. Mutant dynactin in motor neuron disease. *Nat. Genet.*, 33(4):455–456, Apr 2003. 12, 25

S. M. Pulst. Genetic linkage analysis. *Arch. Neurol.*, 56(6):667–672, Jun 1999. 28, 29, 30

S. M. Pulst, A. Nechiporuk, T. Nechiporuk, S. Gispert, X. N. Chen, I. Lopes-Cendes, S. Pearlman, S. Starkman, G. Orozco-Diaz, A. Lunkes, P. DeJong, G. A. Rouleau,

G. Auburger, J. R. Korenberg, C. Figueroa, and S. Sahba. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat. Genet.*, 14(3):269–276, Nov 1996. 283, 305

E. Pupillo, P. Messina, G. Logroscino, E. Beghi, A. Micheli, P. Rosettani, D. Baldini, G. Bianchi, A. Rigamonti, V. Bonito, L. Chiveri, M. Guidotti, M. Rezzonico, S. Vidale, M. Corbo, C. Lunetta, E. Maestri, M. S. Cotelli, M. Filosto, G. Filippini, G. Lauria, G. Mora, L. Papetti, C. Morelli, M. Perini, F. Tavernelli, P. Perrone, M. C. Guaita, D. Testa, F. Sasanelli, A. Galbussera, L. Tremolizzo, C. Ferrarese, A. Galli, E. Vitelli, A. Prelle, N. Riva, M. Ceroni, L. Delodovici, M. Clerici, G. Bono, P. Buzzi, P. Previdi, G. Guarneri, L. Abruzzi, T. Riccardi, L. Lorusso, and L. Mazzini. Long-term survival in amyotrophic lateral sclerosis: a population-based study. *Ann. Neurol.*, 75(2):287–297, Feb 2014. 5

D. Quang, Y. Chen, and X. Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, Mar 2015. 299

B. Quintans, A. Ordonez-Ugalde, P. Cacheiro, A. Carracedo, and M. J. Sobrido. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Appl Transl Genom*, 3(3):60–67, Sep 2014. 296, 297

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/. 63, 65

C. Raczy, R. Petrovski, C. T. Saunders, I. Chorny, S. Kruglyak, E. H. Margulies, H. Y. Chuang, M. Kallberg, S. A. Kumar, A. Liao, K. M. Little, M. P. Stromberg, and S. W. Tanner. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, 29(16):2041–2043, Aug 2013. 48, 225, 288

J. Ramser, M. E. Ahearn, C. Lenski, K. O. Yariz, H. Hellebrand, M. von Rhein, R. D. Clark, R. K. Schmutzler, P. Lichtner, E. P. Hoffman, A. Meindl, and L. Baumbach-Reardon. Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy. *Am. J. Hum. Genet.*, 82(1):188–193, Jan 2008. 154

K. W. Ramsey, T. P. Slavin, G. Graham, G. I. Hirata, V. Balaraman, and L. H. Seaver. Monozygotic twins discordant for trisomy 13. *J Perinatol*, 32(4):306–308, Apr 2012. 34, 271

J. M. Ravits and A. R. La Spada. ALS motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology*, 73(10):805–811, Sep 2009. 4

H. R. Razzaghian, M. H. Shahi, L. A. Forsberg, T. D. de Stahl, D. Absher, N. Dahl, M. P. Westerman, and J. P. Dumanski. Somatic mosaicism for chromosome X and Y aneuploidies in monozygotic twins heterozygous for sickle cell disease mutation. *Am. J. Med. Genet. A*, 152A(10):2595–2598, Oct 2010. 34, 271

A. G. Reaume, J. L. Elliott, E. K. Hoffman, N. W. Kowall, R. J. Ferrante, D. F. Siwek, H. M. Wilcox, D. G. Flood, M. F. Beal, R. H. Brown, R. W. Scott, and W. D. Snider. Motor neurons in Cu/Zn superoxide dismutase-deficient mice develop normally but exhibit enhanced cell death after axonal injury. *Nat. Genet.*, 13(1):43–47, May 1996. 14

E. Reble, C. A. Castellani, M. G. Melka, R. O'Reilly, and S. M. Singh. VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. *Psychiatr. Genet.*, 27(2):62–70, 04 2017. 222

K. Reinert, B. Langmead, D. Weese, and D. J. Evers. Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet*, 16:133–151, 2015. 293

R. T. Relator, A. Terada, and J. Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Med Genomics*, 11(Suppl 2):31, Apr 2018. 304

A. E. Renton, E. Majounie, A. Waite, J. Simon-Sanchez, S. Rollinson, J. R. Gibbs, J. C. Schymick, H. Laaksovirta, J. C. van Swieten, L. Myllykangas, H. Kalimo, A. Paetau, Y. Abramzon, A. M. Remes, A. Kaganovich, S. W. Scholz, J. Duckworth, J. Ding, D. W. Harmer, D. G. Hernandez, J. O. Johnson, K. Mok, M. Ryten, D. Trabzuni, R. J. Guerreiro, R. W. Orrell, J. Neal, A. Murray, J. Pearson, I. E. Jansen, D. Sondervan, H. Seelaar, D. Blake, K. Young, N. Halliwell, J. B. Callister, G. Toulson, A. Richardson, A. Gerhard, J. Snowden, D. Mann, D. Neary, M. A. Nalls, T. Peuralinna, L. Jansson, V. M. Isoviita, A. L. Kaivorinne, M. Holtta-Vuori, E. Ikonen, R. Sulkava, M. Benatar, J. Wuu, A. Chio, G. Restagno, G. Borghero, M. Sabatelli, D. Heckerman, E. Rogaeva, L. Zinman, J. D. Rothstein, M. Sendtner, C. Drepper, E. E. Eichler, C. Alkan, Z. Abdullaev, S. D. Pack, A. Dutra, E. Pak, J. Hardy, A. Singleton, N. M. Williams, P. Heutink, S. Pickering-Brown, H. R. Morris, P. J. Tienari, B. J. Traynor, A. Calvo, S. Cammarosano, C. Moglia, A. Canosa, S. Gallo, M. Brunetti, I. Ossola, G. Mora, K. Marinou, L. Papetti, A. Conte, M. Luigetti,

V. La Bella, R. Spataro, T. Colletti, S. Battistini, F. Giannini, C. Ricci, C. Caponnetto, G. Mancardi, P. Mandich, F. Salvi, I. Bartolomei, J. Mandrioli, P. Sola, M. Corbo, C. Lunetta, S. Penco, M. R. Monsurro, G. Tedeschi, F. L. Conforti, P. Volanti, G. Floris, A. Cannas, V. Piras, M. R. Murru, M. G. Marrosu, M. Pugliatti, A. Ticca, I. Simone, and G. Logroscino. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, 72(2): 257–268, Oct 2011. 11, 12, 20, 22, 38, 271

A. E. Renton, A. Chio, and B. J. Traynor. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.*, 17(1):17–23, Jan 2014. 10, 13, 15, 16, 17, 18, 20, 23, 24, 26, 27, 168, 281

B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, 39(17):e118, Sep 2011. 57

G. Riboldi, C. Zanetta, M. Ranieri, M. Nizzardo, C. Simone, F. Magri, N. Bresolin, G. P. Comi, and S. Corti. Antisense oligonucleotide therapy for the treatment of C9ORF72 ALS/FTD diseases. *Mol. Neurobiol.*, 50(3):721–732, Dec 2014. 7

S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, 17(5):405–424, May 2015. 55, 183, 207, 211, 212, 280, 296, 297, 298, 403

G. M. Ringholz, S. H. Appel, M. Bradshaw, N. A. Cooke, D. M. Mosnik, and P. E. Schulz. Prevalence and patterns of cognitive impairment in sporadic ALS. *Neurology*, 65(4):586–590, Aug 2005. 5

G. R. Ritchie, I. Dunham, E. Zeggini, and P. Flicek. Functional annotation of noncoding sequence variants. *Nat. Methods*, 11(3):294–296, Mar 2014. 299

K. Robasky, N. E. Lewis, and G. M. Church. The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.*, 15(1):56–62, 01 2014. 292

W. Robberecht and T. Philips. The changing scene of amyotrophic lateral sclerosis. *Nat. Rev. Neurosci.*, 14(4):248–264, Apr 2013. 9, 13, 15, 17, 19

S. P. Robertson, Z. A. Jenkins, T. Morgan, L. Ades, S. Aftimos, O. Boute, T. Fiskerstrand, S. Garcia-Minaur, A. Grix, A. Green, V. Der Kaloustian, R. Lewkonia, B. McInnes, M. M. van Haelst, G. Mancini, G. Macini, T. Illes, G. Mortier, R. Newbury-Ecob, L. Nicholson, C. I. Scott, K. Ochman, I. Brozek, D. J. Shears, A. Superti-Furga, M. Suri, M. Whiteford, A. O. Wilkie, and D. Krakow. Frontometaphyseal dysplasia: mutations in FLNA and phenotypic diversity. *Am. J. Med. Genet. A*, 140(16):1726–1736, Aug 2006. 222

J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26, Jan 2011. 52, 67

M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3):511–513, 02 2018. 299

D. Ronchi, G. Riboldi, R. Del Bo, N. Ticozzi, M. Scarlato, D. Galimberti, S. Corti, V. Silani, N. Bresolin, and G. P. Comi. CHCHD10 mutations in Italian patients with sporadic amyotrophic lateral sclerosis. *Brain*, 138(Pt 8):e372, Aug 2015. 106, 107

D. R. Rosen. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 364(6435):362, Jul 1993. 11, 12, 13, 34, 168

C. Rothenberg, D. Srinivasan, L. Mah, S. Kaushik, C. M. Peterhoff, J. Ugolino, S. Fang, A. M. Cuervo, R. A. Nixon, and M. J. Monteiro. Ubiquilin functions in autophagy and is degraded by chaperone-mediated autophagy. *Hum. Mol. Genet.*, 19(16):3219–3232, Aug 2010. 24

D. M. Ruddy, M. J. Parton, A. Al-Chalabi, C. M. Lewis, C. Vance, B. N. Smith, P. N. Leigh, J. F. Powell, T. Siddique, E. P. Meyjes, F. Baas, V. de Jong, and C. E. Shaw. Two families with familial amyotrophic lateral sclerosis are linked to a novel locus on chromosome 16q. *Am. J. Hum. Genet.*, 73(2):390–396, Aug 2003. 17

R. A. Saccon, R. K. Bunton-Stasyshyn, E. M. Fisher, and P. Fratta. Is SOD1 loss of function involved in amyotrophic lateral sclerosis? *Brain*, 136(Pt 8):2342–2358, Aug 2013. 14

S. L. Sawyer, T. Hartley, D. A. Dyment, C. L. Beaulieu, J. Schwartzentruber, A. Smith, H. M. Bedford, G. Bernard, F. P. Bernier, B. Brais, D. E. Bulman, J. Warman Chardon, D. Chitayat, J. Deladoey, B. A. Fernandez, P. Frosk, M. T. Geraghty,

B. Gerull, W. Gibson, R. M. Gow, G. E. Graham, J. S. Green, E. Heon, G. Horvath, A. M. Innes, N. Jabado, R. H. Kim, R. K. Koenekoop, A. Khan, O. J. Lehmann, R. Mendoza-Londono, J. L. Michaud, S. M. Nikkel, L. S. Penney, C. Polychronakos, J. Richer, G. A. Rouleau, M. E. Samuels, V. M. Siu, O. Suchowersky, M. A. Tarnopolsky, G. Yoon, F. R. Zahir, J. Majewski, and K. M. Boycott. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin. Genet.*, 89(3):275–284, Mar 2016. 32

B. Schmid, A. Hruscha, S. Hogl, J. Banzhaf-Strathmann, K. Strecker, J. van der Zee, M. Teucke, S. Eimer, J. Hegermann, M. Kittelmann, E. Kremmer, M. Cruts, B. Solchenberger, L. Hasenkamp, F. van Bebber, C. Van Broeckhoven, D. Edbauer, S. F. Lichtenthaler, and C. Haass. Loss of ALS-associated TDP-43 in zebrafish causes muscle degeneration, vascular dysfunction, and reduced motor neuron axon outgrowth. *Proc. Natl. Acad. Sci. U.S.A.*, 110(13):4986–4991, Mar 2013. 16

M. Schwartz. *WriteXLS: Cross-Platform Perl Based R Function to Create Excel 2003 (XLS) and Excel 2007 (XLSX) Files*, 2015. URL https://CRAN.R-project.org/package=WriteXLS. 65

J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, 11(4):361–362, Apr 2014. 57, 216

B. M. Schwenk, H. Hartmann, A. Serdaroglu, M. H. Schludi, D. Hornburg, F. Meissner, D. Orozco, A. Colombo, S. Tahirovic, M. Michaelsen, F. Schreiber, S. Haupt, M. Peitz, O. Brustle, C. Kupper, T. Klopstock, M. Otto, A. C. Ludolph, T. Arzberger, P. H. Kuhn, and D. Edbauer. TDP-43 loss of function inhibits endosomal trafficking and alters trophic signaling in neurons. *EMBO J.*, 35(21):2350–2370, 11 2016. 16

E. G. Seaby, R. J. Pengelly, and S. Ennis. Exome sequencing explained: a practical guide to its clinical application. *Brief Funct Genomics*, 15(5):374–384, Sep 2016. 45

A. Sharma, A. K. Lyashchenko, L. Lu, S. E. Nasrabady, M. Elmaleh, M. Mendelsohn, A. Nemes, J. C. Tapia, G. Z. Mentis, and N. A. Shneider. ALS-associated mutant FUS induces selective motor neuron degeneration through toxic gain of function. *Nat Commun*, 7:10465, Feb 2016. 18

A. Shatunov, K. Mok, S. Newhouse, M. E. Weale, B. Smith, C. Vance, L. Johnson, J. H. Veldink, M. A. van Es, L. H. van den Berg, W. Robberecht, P. Van Damme,

O. Hardiman, A. E. Farmer, C. M. Lewis, A. W. Butler, O. Abel, P. M. Andersen, I. Fogh, V. Silani, A. Chio, B. J. Traynor, J. Melki, V. Meininger, J. E. Landers, P. McGuffin, J. D. Glass, H. Pall, P. N. Leigh, J. Hardy, R. H. Brown, J. F. Powell, R. W. Orrell, K. E. Morrison, P. J. Shaw, C. E. Shaw, and A. Al-Chalabi. Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurol*, 9(10):986–994, Oct 2010. 19, 27, 281

S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29 (1):308–311, Jan 2001. 208

N. Shibata, A. Hirano, M. Kobayashi, T. Siddique, H. X. Deng, W. Y. Hung, T. Kato, and K. Asayama. Intense superoxide dismutase-1 immunoreactivity in intracytoplasmic hyaline inclusions of familial amyotrophic lateral sclerosis with posterior column involvement. *J. Neuropathol. Exp. Neurol.*, 55(4):481–490, Apr 1996. 15, 168

H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. Day, T. R. Gaunt, and C. Campbell. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–1543, May 2015. 299

T. Siddique, D. A. Figlewicz, M. A. Pericak-Vance, J. L. Haines, G. Rouleau, A. J. Jeffers, P. Sapp, W. Y. Hung, J. Bebout, and D. McKenna-Yasek. Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N. Engl. J. Med.*, 324(20):1381–1384, May 1991. 13

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, Aug 2005. 55, 57, 212

F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7:539, 2011. 55, 57, 195

C. L. Simpson, R. Lemmens, K. Miskiewicz, W. J. Broom, V. K. Hansen, P. W. van Vught, J. E. Landers, P. Sapp, L. Van Den Bosch, J. Knight, B. M. Neale, M. R.

Turner, J. H. Veldink, R. A. Ophoff, V. B. Tripathi, A. Beleza, M. N. Shah, P. Proitsi, A. Van Hoecke, P. Carmeliet, H. R. Horvitz, P. N. Leigh, C. E. Shaw, L. H. van den Berg, P. C. Sham, J. F. Powell, P. Verstreken, R. H. Brown, W. Robberecht, and A. Al-Chalabi. Variants of the elongator protein 3 (ELP3) gene are associated with motor neuron degeneration. *Hum. Mol. Genet.*, 18(3):472–481, Feb 2009. 12

R. Sivadasan, D. Hornburg, C. Drepper, N. Frank, S. Jablonka, A. Hansel, X. Lojewski, J. Sterneckert, A. Hermann, P. J. Shaw, P. G. Ince, M. Mann, F. Meissner, and M. Sendtner. C9ORF72 interaction with cofilin modulates actin dynamics in motor neurons. *Nat. Neurosci.*, 19(12):1610–1618, 12 2016. 21

B. N. Smith, N. Ticozzi, C. Fallini, A. S. Gkazi, S. Topp, K. P. Kenna, E. L. Scotter, J. Kost, P. Keagle, J. W. Miller, D. Calini, C. Vance, E. W. Danielson, C. Troakes, C. Tiloca, S. Al-Sarraj, E. A. Lewis, A. King, C. Colombrita, V. Pensato, B. Castellotti, J. de Belleroche, F. Baas, A. L. ten Asbroek, P. C. Sapp, D. McKenna-Yasek, R. L. McLaughlin, M. Polak, S. Asress, J. Esteban-Perez, J. L. Munoz-Blanco, M. Simpson, W. van Rheenen, F. P. Diekstra, G. Lauria, S. Duga, S. Corti, C. Cereda, L. Corrado, G. Soraru, K. E. Morrison, K. L. Williams, G. A. Nicholson, I. P. Blair, P. A. Dion, C. S. Leblond, G. A. Rouleau, O. Hardiman, J. H. Veldink, L. H. van den Berg, A. Al-Chalabi, H. Pall, P. J. Shaw, M. R. Turner, K. Talbot, F. Taroni, A. Garcia-Redondo, Z. Wu, J. D. Glass, C. Gellera, A. Ratti, R. H. Brown, V. Silani, C. E. Shaw, J. E. Landers, S. D'Alfonso, L. Mazzini, G. P. Comi, R. Del Bo, M. Ceroni, S. Gagliardi, G. Querin, and C. Bertolin. Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron*, 84(2):324–331, Oct 2014. 12, 25, 28, 35, 282

J. A. Solski, S. Yang, G. A. Nicholson, N. Luquin, K. L. Williams, R. Fernando, R. Pamphlett, and I. P. Blair. A novel TARDBP insertion/deletion mutation in the flail arm variant of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 13(5): 465–470, Sep 2012. 15

J. Sreedharan and R. H. Brown. Amyotrophic lateral sclerosis: Problems and prospects. *Ann. Neurol.*, 74(3):309–316, Sep 2013. 13

J. Sreedharan, I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, 319(5870):1668–1672, Mar 2008. 11, 12, 15, 34, 148, 167, 168, 205, 278

H. Stranneheim and A. Wedell. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *J. Intern. Med.*, 279(1):3–15, Jan 2016. 32

S. Sun, S. C. Ling, J. Qiu, C. P. Albuquerque, Y. Zhou, S. Tokunaga, H. Li, H. Qiu, A. Bui, G. W. Yeo, E. J. Huang, K. Eggan, H. Zhou, X. D. Fu, C. Lagier-Tourenne, and D. W. Cleveland. ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nat Commun*, 6:6171, Jan 2015. 19

A. Swaminathan, M. Bouffard, M. Liao, S. Ryan, J. Bennion Callister, S. M. Pickering-Brown, G. A. B. Armstrong, and P. Drapeau. Expression of C9orf72-related dipeptides impairs motor function in a vertebrate model. *Hum. Mol. Genet.*, Mar 2018. 22

B. Swinnen and W. Robberecht. The phenotypic variability of amyotrophic lateral sclerosis. *Nat Rev Neurol*, 10(11):661–670, Nov 2014. 3, 4, 5

M. Synofzik, W. Maetzler, T. Grehl, J. Prudlo, J. M. Vom Hagen, T. Haack, P. Rebassoo, M. Munz, L. Schols, and S. Biskup. Screening in ALS and FTD patients reveals 3 novel UBQLN2 mutations outside the PXX domain and a pure FTD phenotype. *Neurobiol. Aging*, 33(12):13–17, Dec 2012. 23

D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(Database issue):D447–452, Jan 2015. 57

Y. Takahashi, Y. Fukuda, J. Yoshimura, A. Toyoda, K. Kurppa, H. Moritoyo, V. V. Belzil, P. A. Dion, K. Higasa, K. Doi, H. Ishiura, J. Mitsui, H. Date, B. Ahsan, T. Matsukawa, Y. Ichikawa, T. Moritoyo, M. Ikoma, T. Hashimoto, F. Kimura, S. Murayama, O. Onodera, M. Nishizawa, M. Yoshida, N. Atsuta, G. Sobue, J. A. Fifita, K. L. Williams, I. P. Blair, G. A. Nicholson, P. Gonzalez-Perez, R. H. Brown, M. Nomoto, K. Elenius, G. A. Rouleau, A. Fujiyama, S. Morishita, J. Goto, S. Tsuji, R. Nakamura, H. Watanabe, Y. Izumi, R. Kaji, M. Morita, K. Ogaki, A. Taniguchi, I. Aiba, K. Mizoguchi, K. Okamoto, K. Hasegawa, M. Aoki, A. Kawata, I. Nakano, K. Abe, M. Oda, M. Konagaya, T. Imai, M. Nakagawa, T. Fujita, H. Sasaki, and M. Nishizawa. ERBB4 mutations that disrupt the neuregulin-ErbB4 pathway cause amyotrophic lateral sclerosis type 19. *Am. J. Hum. Genet.*, 93(5):900–905, Nov 2013. 12, 25

K. Talbot. Motor neuron disease: the bare essentials. *Pract Neurol*, 9(5):303–309, Oct 2009. 5

A. Y. Tan and J. L. Manley. The TET family of proteins: functions and roles in disease. *J Mol Cell Biol*, 1(2):82–92, Dec 2009. 18

C. F. Tan, H. Eguchi, A. Tagawa, O. Onodera, T. Iwasaki, A. Tsujino, M. Nishizawa, A. Kakita, and H. Takahashi. TDP-43 immunoreactivity in neuronal inclusions in familial amyotrophic lateral sclerosis with or without SOD1 gene mutation. *Acta Neuropathol.*, 113(5):535–542, May 2007. 15

K. Tanaka and N. Matsuda. Proteostasis and neurodegeneration: the roles of proteasomal degradation and autophagy. *Biochim. Biophys. Acta*, 1843(1):197–204, Jan 2014. 10

J. P. Taylor, R. H. Brown, and D. W. Cleveland. Decoding ALS: from genes to mechanism. *Nature*, 539(7628):197–206, 11 2016. 7, 8, 11, 13, 21, 26

J. K. Teer and J. C. Mullikin. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, 19(R2):R145–151, Oct 2010. 45

A. Telenti, L. C. Pierce, W. H. Biggs, J. di Iulio, E. H. Wong, M. M. Fabani, E. F. Kirkness, A. Moustafa, N. Shah, C. Xie, S. C. Brewerton, N. Bulsara, C. Garner, G. Metzker, E. Sandoval, B. A. Perkins, F. J. Och, Y. Turpaz, and J. C. Venter. Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 113(42): 11901–11906, 10 2016. 269, 290, 291, 296

A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(32):12996–13001, Aug 2013. 304

A. Terada, R. Yamada, K. Tsuda, and J. Sese. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics*, 32(22):3513–3515, 11 2016. 304

E. Teyssou, L. Chartier, M. Albert, A. Bouscary, J. C. Antoine, J. P. Camdessanche, F. Rotolo, P. Couratier, F. Salachas, D. Seilhean, and S. Millecamps. Genetic analysis of CHCHD10 in French familial amyotrophic lateral sclerosis patients. *Neurobiol. Aging*, 42:1–3, 06 2016. 106

M. Therrien, P. A. Dion, and G. A. Rouleau. ALS: Recent Developments from Genetics Studies. *Curr Neurol Neurosci Rep*, 16(6):59, Jun 2016. 8, 10, 15, 280

J. Theuns, A. Verstraeten, K. Sleegers, E. Wauters, I. Gijselinck, S. Smolders, D. Crosiers, E. Corsmit, E. Elinck, M. Sharma, R. Kruger, S. Lesage, A. Brice, S. J. Chung, M. J. Kim, Y. J. Kim, O. A. Ross, Z. K. Wszolek, E. Rogaeva, Z. Xi, A. E. Lang, C. Klein, A. Weissbach, G. D. Mellick, P. A. Silburn, G. M. Hadjigeorgiou, E. Dardiotis, N. Hattori, K. Ogaki, E. K. Tan, Y. Zhao, J. Aasly, E. M. Valente, S. Petrucci, G. Annesi, A. Quattrone, C. Ferrarese, L. Brighina, A. Deutschlander, A. Puschmann, C. Nilsson, G. Garraux, M. S. LeDoux, R. F. Pfeiffer, M. Boczarska-Jedynak, G. Opala, D. M. Maraganore, S. Engelborghs, P. P. De Deyn, P. Cras, M. Cruts, C. Van Broeckhoven, D. M. Maraganore, M. J. Farrer, J. O. Aasly, R. Kruger, A. Elbaz, J. P. Ioannidis, G. Annesi, E. M. Valente, M. Bozi, A. Brice, M. Curie, A. Carmine-Belin, J. Carr, C. Carroll, S. D. Chen, S. J. Chung, C. Cosentino, S. Cresswell, A. Deutschlaender, C. Ferrarese, T. Foroud, G. Garraux, S. Goldwurm, G. Hadjigeorgiou, M. C. Chartier-Harlin, A. Hassan, N. Hattori, F. Hentati, B. S. Jeon, H. Kawakami, Y. J. Kim, A. Kishore, C. Klein, S. Koks, D. Krainc, R. Kruger, A. Krygowska-Wajs, J. J. Lin, T. Lynch, D. M. Maraganore, G. Mellick, K. E. Morrison, R. P. Munhoz, G. Opala, P. Pastor, H. Payami, S. N. Pchelina, S. Petersburg, M. S. Petersen, A. Puschmann, B. Ritz, E. Rogaeva, A. Sazci, J. Slawek, L. Stefanis, E. K. Tan, T. Toda, M. Toft, C. Van Broeckhoven, K. Wirdefeldt, D. Woitalla, Z. K. Wszolek, and A. Zimprich. Global investigation and meta-analysis of the C9orf72 (G4C2)n repeat in Parkinson disease. *Neurology*, 83(21):1906–1913, Nov 2014. 20

J. Thusberg, A. Olatubosun, and M. Vihinen. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, 32(4):358–368, Apr 2011. 211, 212

S. Tian, H. Yan, C. Neuhauser, and S. L. Slager. An analytical workflow for accurate variant discovery in highly divergent regions. *BMC Genomics*, 17:703, 09 2016. 268, 293

N. Ticozzi, C. Vance, A. L. Leclerc, P. Keagle, J. D. Glass, D. McKenna-Yasek, P. C. Sapp, V. Silani, D. A. Bosco, C. E. Shaw, R. H. Brown, and J. E. Landers. Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 156B (3):285–290, Apr 2011. 9

N. J. Timpson, C. M. T. Greenwood, N. Soranzo, D. J. Lawson, and J. B. Richards. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.*, 19(2):110–124, Feb 2018. 32

E. Tiryaki and H. A. Horak. ALS and other motor neuron diseases. *Continuum (Minneap Minn)*, 20(5 Peripheral Nervous System Disorders):1185–1207, Oct 2014. 2, 3, 4, 5

A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.*, 19(9):581–590, Sep 2018. 281

F. Trojsi, M. R. Monsurro, and G. Tedeschi. Exposure to environmental toxicants and pathogenesis of amyotrophic lateral sclerosis: state of the art and research perspectives. *Int J Mol Sci*, 14(8):15286–15311, Jul 2013. 6

M. R. Turner and M. C. Kiernan. The standard of care in amyotrophic lateral sclerosis: a centralised multidisciplinary clinic encounter sets a new benchmark for a uniquely challenging neurodegenerative disorder. *J. Neurol. Neurosurg. Psychiatry*, 86(5): 481–482, May 2015. 7

P. N. Valdmanis, N. Dupre, J. P. Bouchard, W. Camu, F. Salachas, V. Meininger, M. Strong, and G. A. Rouleau. Three families with amyotrophic lateral sclerosis and frontotemporal dementia with evidence of linkage to chromosome 9p. *Arch. Neurol.*, 64(2):240–245, Feb 2007. 19

E. M. Valente, P. M. Abou-Sleiman, V. Caputo, M. M. Muqit, K. Harvey, S. Gispert, Z. Ali, D. Del Turco, A. R. Bentivoglio, D. G. Healy, A. Albanese, R. Nussbaum, R. Gonzalez-Maldonado, T. Deller, S. Salvi, P. Cortelli, W. P. Gilks, D. S. Latchman, R. J. Harvey, B. Dallapiccola, G. Auburger, and N. W. Wood. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science*, 304(5674):1158–1160, May 2004. 154

M. van Blitterswijk, M. A. van Es, E. A. Hennekam, D. Dooijes, W. van Rheenen, J. Medic, P. R. Bourque, H. J. Schelhaas, A. J. van der Kooi, M. de Visser, P. I. de Bakker, J. H. Veldink, and L. H. van den Berg. Evidence for an oligogenic basis of amyotrophic lateral sclerosis. *Hum. Mol. Genet.*, 21(17):3776–3784, Sep 2012a. 278, 279, 280

M. van Blitterswijk, L. Vlam, M. A. van Es, W. L. van der Pol, E. A. Hennekam, D. Dooijes, H. J. Schelhaas, A. J. van der Kooi, M. de Visser, J. H. Veldink, and L. H. van den Berg. Genetic overlap between apparently sporadic motor neuron diseases. *PLoS ONE*, 7(11):e48983, 2012b. 10

P. Van Damme, W. Robberecht, and L. Van Den Bosch. Modelling amyotrophic lateral sclerosis: progress and possibilities. *Dis Model Mech*, 10(5):537–549, 05 2017. 303

V. M. Van Deerlin, J. B. Leverenz, L. M. Bekris, T. D. Bird, W. Yuan, L. B. Elman, D. Clay, E. M. Wood, A. S. Chen-Plotkin, M. Martinez-Lage, E. Steinbart, L. McCluskey, M. Grossman, M. Neumann, I. L. Wu, W. S. Yang, R. Kalb, D. R. Galasko, T. J. Montine, J. Q. Trojanowski, V. M. Lee, G. D. Schellenberg, and C. E. Yu. TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurol*, 7(5):409–416, May 2008. 17

V. M. Van Deerlin, P. M. Sleiman, M. Martinez-Lage, A. Chen-Plotkin, L. S. Wang, N. R. Graff-Radford, D. W. Dickson, R. Rademakers, B. F. Boeve, M. Grossman, S. E. Arnold, D. M. Mann, S. M. Pickering-Brown, H. Seelaar, P. Heutink, J. C. van Swieten, J. R. Murrell, B. Ghetti, S. Spina, J. Grafman, J. Hodges, M. G. Spillantini, S. Gilman, A. P. Lieberman, J. A. Kaye, R. L. Woltjer, E. H. Bigio, M. Mesulam, S. Al-Sarraj, C. Troakes, R. N. Rosenberg, C. L. White, I. Ferrer, A. Llado, M. Neumann, H. A. Kretzschmar, C. M. Hulette, K. A. Welsh-Bohmer, B. L. Miller, A. Alzualde, A. Lopez de Munain, A. C. McKee, M. Gearing, A. I. Levey, J. J. Lah, J. Hardy, J. D. Rohrer, T. Lashley, I. R. Mackenzie, H. H. Feldman, R. L. Hamilton, S. T. Dekosky, J. van der Zee, S. Kumar-Singh, C. Van Broeckhoven, R. Mayeux, J. P. Vonsattel, J. C. Troncoso, J. J. Kril, J. B. Kwok, G. M. Halliday, T. D. Bird, P. G. Ince, P. J. Shaw, N. J. Cairns, J. C. Morris, C. A. McLean, C. DeCarli, W. G. Ellis, S. H. Freeman, M. P. Frosch, J. H. Growdon, D. P. Perl, M. Sano, D. A. Bennett, J. A. Schneider, T. G. Beach, E. M. Reiman, B. K. Woodruff, J. Cummings, H. V. Vinters, C. A. Miller, H. C. Chui, I. Alafuzoff, P. Hartikainen, D. Seilhean, D. Galasko, E. Masliah, C. W. Cotman, M. T. Tunon, M. C. Martinez, D. G. Munoz, S. L. Carroll, D. Marson, P. F. Riederer, N. Bogdanovic, G. D. Schellenberg, H. Hakonarson, J. Q. Trojanowski, and V. M. Lee. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.*, 42(3):234–239, Mar 2010. 19

J. van der Zee, I. Gijselinck, L. Dillen, T. Van Langenhove, J. Theuns, S. Engelborghs, S. Philtjens, M. Vandenbulcke, K. Sleegers, A. Sieben, V. Baumer, G. Maes, E. Corsmit, B. Borroni, A. Padovani, S. Archetti, R. Perneczky, J. Diehl-Schmid, A. de Mendonca, G. Miltenberger-Miltenyi, S. Pereira, J. Pimentel, B. Nacmias, S. Bagnoli, S. Sorbi, C. Graff, H. H. Chiang, M. Westerlund, R. Sanchez-Valle, A. Llado, E. Gelpi, I. Santana, M. R. Almeida, B. Santiago, G. Frisoni, O. Zanetti, C. Bonvicini, M. Synofzik, W. Maetzler, J. M. Vom Hagen, L. Schols, M. T. Heneka, F. Jessen, R. Matej, E. Parobkova, G. G. Kovacs, T. Strobel, S. Sarafov, I. Tournev, A. Jordanova, A. Danek, T. Arzberger, G. M. Fabrizi, S. Testi, E. Salmon, P. Santens,

J. J. Martin, P. Cras, R. Vandenberghe, P. P. De Deyn, M. Cruts, C. Van Broeckhoven, J. van der Zee, I. Gijselinck, L. Dillen, T. Van Langenhove, J. Theuns, S. Philtjens, K. Sleegers, V. Baumer, G. Maes, E. Corsmit, M. Cruts, C. Van Broeckhoven, J. van der Zee, I. Gijselinck, L. Dillen, T. Van Langenhove, S. Philtjens, J. Theuns, K. Sleegers, V. Baumer, G. Maes, M. Cruts, C. Van Broeckhoven, S. Engelborghs, P. P. De Deyn, P. Cras, S. Engelborghs, P. P. De Deyn, M. Vandenbulcke, M. Vandenbulcke, B. Borroni, A. Padovani, S. Archetti, R. Perneczky, J. Diehl-Schmid, M. Synofzik, W. Maetzler, J. Muller Vom Hagen, L. Schols, M. Synofzik, W. Maetzler, J. Muller Vom Hagen, L. Schols, M. T. Heneka, F. Jessen, A. Ramirez, D. Kurzwelly, C. Sachtleben, W. Mairer, A. de Mendonca, G. Miltenberger-Miltenyi, S. Pereira, C. Firmo, J. Pimentel, R. Sanchez-Valle, A. Llado, A. Antonell, J. Molinuevo, E. Gelpi, C. Graff, H. H. Chiang, M. Westerlund, C. Graff, A. Kinhult Stahlbom, H. Thonberg, I. Nennesmo, A. Borjesson-Hanson, B. Nacmias, S. Bagnoli, S. Sorbi, V. Bessi, I. Piaceri, I. Santana, B. Santiago, I. Santana, M. Helena Ribeiro, M. Rosario Almeida, C. Oliveira, J. Massano, C. Garret, P. Pires, G. Frisoni, O. Zanetti, C. Bonvicini, S. Sarafov, I. Tournev, A. Jordanova, I. Tournev, G. G. Kovacs, T. Strobel, M. T. Heneka, F. Jessen, A. Ramirez, D. Kurzwelly, C. Sachtleben, W. Mairer, F. Jessen, R. Matej, E. Parobkova, A. Danel, T. Arzberger, G. Maria Fabrizi, S. Testi, S. Ferrari, T. Cavallaro, E. Salmon, P. Santens, and P. Cras. A pan-European study of the C9orf72 repeat associated with FTLD: geographic prevalence, genomic instability, and intermediate repeats. *Hum. Mutat.*, 34(2):363–373, Feb 2013. 20

E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.*, 322(1):12–20, Mar 2014. 45

J. van Dongen, P. E. Slagboom, H. H. Draisma, N. G. Martin, and D. I. Boomsma. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.*, 13(9):640–653, Sep 2012. 222

P. T. van Doormaal, W. van Rheenen, M. van Blitterswijk, R. D. Schellevis, H. J. Schelhaas, M. de Visser, A. J. van der Kooi, J. H. Veldink, and L. H. van den Berg. UBQLN2 in familial amyotrophic lateral sclerosis in The Netherlands. *Neurobiol. Aging*, 33(9):7–2233, Sep 2012. 23

R. P. A. van Eijk, A. R. Jones, W. Sproviero, A. Shatunov, P. J. Shaw, P. N. Leigh, C. A. Young, C. E. Shaw, G. Mora, J. Mandrioli, G. Borghero, P. Volanti, F. P. Diekstra, W. van Rheenen, E. Verstraete, M. J. C. Eijkemans, J. H. Veldink, A. Chio,

A. Al-Chalabi, L. H. van den Berg, M. A. van Es, C. Allen, C. Counsell, A. Farrin, A. Al-Chalabi, B. Dickie, J. Kelly, P. N. Leigh, C. L. Murphy, C. Payan, G. Reynolds, P. Shaw, I. N. Steen, M. Thornhill, J. Waters, J. Zajicek, P. N. Leigh, A. Al-Chalabi, P. J. Shaw, C. A. Young, M. Thornhill, I. N. Steen, C. L. Murphy, K. E. Morrison, S. Dhariwal, R. Hornabrook, L. Savage, D. J. Burn, T. K. Khoo, J. Kelly, C. L. Murphy, A. Al-Chalabi, A. Dougherty, P. N. Leigh, L. Wijesekera, M. Thornhill, C. M. Ellis, R. Ali, K. O'Hanlon, J. Panicker, L. Pate, P. Ray, L. Wyatt, C. A. Young, L. Copeland, J. Ealing, H. Hamdalla, I. Leroi, C. Murphy, F. O'Keeffe, E. Oughton, L. Partington, P. Paterson, D. Rog, A. Sathish, D. Sexton, J. Smith, H. Vanek, S. Dodds, T. L. Williams, I. N. Steen, J. Clarke, C. Eziefula, R. Howard, R. Orrell, K. Sidle, R. Sylvester, W. Barrett, C. Merritt, K. Talbot, M. R. Turner, C. Whatley, C. Williams, J. Williams, C. Cosby, C. O. Hanemann, I. Imam, C. Phillips, L. Timings, S. E. Crawford, C. Hewamadduma, R. Hibberd, H. Hollinger, C. McDermott, G. Mills, M. Rafiq, P. J. Shaw, A. Taylor, E. Waines, T. Walsh, R. Addison-Jones, J. Birt, M. Hare, T. Majid, R. Tortelli, E. D'Errico, I. Bartolomei, E. Barbarossa, B. Depau, E. Costantino, E. D'Amico, A. Uncini, C. Manzoli, R. Quatrale, E. Sette, E. Montanari, M. Merello, D. Zarcone, M. Mascolo, M. Vignolo, S. Messina, C. Morelli, K. Marinou, L. Papetti, C. Lunetta, K. Gorni, D. De Cicco, C. Pipia, P. Sola, E. Georgoulopoulou, A. Sagnelli, G. Tedeschi, G. Oggioni, N. Nasuelli, C. D'Ascenzo, V. Cima, M. Aiello, R. Rizzi, E. Rinaldi, M. Luigetti, A. Conte, A. Torzini, G. Greco, R. Mutani, G. Fuda, and M. A. Tommasi. Meta-analysis of pharmacogenetic interactions in amyotrophic lateral sclerosis clinical trials. *Neurology*, 89(18):1915–1922, Oct 2017. 144

M. A. van Es, J. H. Veldink, C. G. Saris, H. M. Blauw, P. W. van Vught, A. Birve, R. Lemmens, H. J. Schelhaas, E. J. Groen, M. H. Huisman, A. J. van der Kooi, M. de Visser, C. Dahlberg, K. Estrada, F. Rivadeneira, A. Hofman, M. J. Zwarts, P. T. van Doormaal, D. Rujescu, E. Strengman, I. Giegling, P. Muglia, B. Tomik, A. Slowik, A. G. Uitterlinden, C. Hendrich, S. Waibel, T. Meyer, A. C. Ludolph, J. D. Glass, S. Purcell, S. Cichon, M. M. Nothen, H. E. Wichmann, S. Schreiber, S. H. Vermeulen, L. A. Kiemeney, J. H. Wokke, S. Cronin, R. L. McLaughlin, O. Hardiman, K. Fumoto, R. J. Pasterkamp, V. Meininger, J. Melki, P. N. Leigh, C. E. Shaw, J. E. Landers, A. Al-Chalabi, R. H. Brown, W. Robberecht, P. M. Andersen, R. A. Ophoff, and L. H. van den Berg. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nat. Genet.*, 41(10):1083–1087, Oct 2009. 12, 19, 27, 281

M. A. van Es, C. Dahlberg, A. Birve, J. H. Veldink, L. H. van den Berg, and P. M.

Andersen. Large-scale SOD1 mutation screening provides evidence for genetic heterogeneity in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatr.*, 81(5): 562–566, May 2010. 14

A. Van Hoecke, L. Schoonaert, R. Lemmens, M. Timmers, K. A. Staats, A. S. Laird, E. Peeters, T. Philips, A. Goris, B. Dubois, P. M. Andersen, A. Al-Chalabi, V. Thijs, A. M. Turnley, P. W. van Vught, J. H. Veldink, O. Hardiman, L. Van Den Bosch, P. Gonzalez-Perez, P. Van Damme, R. H. Brown, L. H. van den Berg, and W. Robberecht. EPHA4 is a disease modifier of amyotrophic lateral sclerosis in animal models and in humans. *Nat. Med.*, 18(9):1418–1422, Sep 2012. 27

W. van Rheenen, M. van Blitterswijk, M. H. Huisman, L. Vlam, P. T. van Doormaal, M. Seelen, J. Medic, D. Dooijes, M. de Visser, A. J. van der Kooi, J. Raaphorst, H. J. Schelhaas, W. L. van der Pol, J. H. Veldink, and L. H. van den Berg. Hexanucleotide repeat expansions in C9ORF72 in the spectrum of motor neuron diseases. *Neurology*, 79(9):878–882, Aug 2012. 20

W. van Rheenen, A. Shatunov, A. M. Dekker, R. L. McLaughlin, F. P. Diekstra, S. L. Pulit, R. A. van der Spek, U. Vosa, S. de Jong, M. R. Robinson, J. Yang, I. Fogh, P. T. van Doormaal, G. H. Tazelaar, M. Koppers, A. M. Blokhuis, W. Sproviero, A. R. Jones, K. P. Kenna, K. R. van Eijk, O. Harschnitz, R. D. Schellevis, W. J. Brands, J. Medic, A. Menelaou, A. Vajda, N. Ticozzi, K. Lin, B. Rogelj, K. Vrabec, M. Ravnik-Glavac, B. Koritnik, J. Zidar, L. Leonardis, L. D. Groselj, S. Millecamps, F. Salachas, V. Meininger, M. de Carvalho, S. Pinto, J. S. Mora, R. Rojas-Garcia, M. Polak, S. Chandran, S. Colville, R. Swingler, K. E. Morrison, P. J. Shaw, J. Hardy, R. W. Orrell, A. Pittman, K. Sidle, P. Fratta, A. Malaspina, S. Topp, S. Petri, S. Abdulla, C. Drepper, M. Sendtner, T. Meyer, R. A. Ophoff, K. A. Staats, M. Wiedau-Pazos, C. Lomen-Hoerth, V. M. Van Deerlin, J. Q. Trojanowski, L. Elman, L. McCluskey, A. N. Basak, C. Tunca, H. Hamzeiy, Y. Parman, T. Meitinger, P. Lichtner, M. Radivojkov-Blagojevic, C. R. Andres, C. Maurel, G. Bensimon, B. Landwehrmeyer, A. Brice, C. A. Payan, S. Saker-Delye, A. Durr, N. W. Wood, L. Tittmann, W. Lieb, A. Franke, M. Rietschel, S. Cichon, M. M. Nothen, P. Amouyel, C. Tzourio, J. F. Dartigues, A. G. Uitterlinden, F. Rivadeneira, K. Estrada, A. Hofman, C. Curtis, H. M. Blauw, A. J. van der Kooi, M. de Visser, A. Goris, M. Weber, C. E. Shaw, B. N. Smith, O. Pansarasa, C. Cereda, R. Del Bo, G. P. Comi, S. D'Alfonso, C. Bertolin, G. Soraru, L. Mazzini, V. Pensato, C. Gellera, C. Tiloca, A. Ratti, A. Calvo, C. Moglia, M. Brunetti, S. Arcuti, R. Capozzo, C. Zecca, C. Lunetta, S. Penco, N. Riva, A. Padovani, M. Filosto, B. Muller, R. J. Stuit, I. Blair, K. Zhang, E. P. McCann, J. A. Fifita, G. A. Nicholson, D. B. Rowe,

R. Pamphlett, M. C. Kiernan, J. Grosskreutz, O. W. Witte, T. Ringer, T. Prell, B. Stubendorff, I. Kurth, C. A. Hubner, P. N. Leigh, F. Casale, A. Chio, E. Beghi, E. Pupillo, R. Tortelli, G. Logroscino, J. Powell, A. C. Ludolph, J. H. Weishaupt, W. Robberecht, P. Van Damme, L. Franke, T. H. Pers, R. H. Brown, J. D. Glass, J. E. Landers, O. Hardiman, P. M. Andersen, P. Corcia, P. Vourc'h, V. Silani, N. R. Wray, P. M. Visscher, P. I. de Bakker, M. A. van Es, R. J. Pasterkamp, C. M. Lewis, G. Breen, A. Al-Chalabi, L. H. van den Berg, and J. H. Veldink. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.*, 48(9):1043–1048, 09 2016. 12, 27, 35, 154, 281

C. Vance, A. Al-Chalabi, D. Ruddy, B. N. Smith, X. Hu, J. Sreedharan, T. Siddique, H. J. Schelhaas, B. Kusters, D. Troost, F. Baas, V. de Jong, and C. E. Shaw. Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain*, 129(Pt 4):868–876, Apr 2006. 19

C. Vance, B. Rogelj, T. Hortobagyi, K. J. De Vos, A. L. Nishimura, J. Sreedharan, X. Hu, B. Smith, D. Ruddy, P. Wright, J. Ganesalingam, K. L. Williams, V. Tripathi, S. Al-Saraj, A. Al-Chalabi, P. N. Leigh, I. P. Blair, G. Nicholson, J. de Belleroche, J. M. Gallo, C. C. Miller, and C. E. Shaw. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science*, 323(5918):1208–1211, Feb 2009. 11, 12, 17, 19, 34, 148, 278

C. Vance, E. L. Scotter, A. L. Nishimura, C. Troakes, J. C. Mitchell, C. Kathe, H. Urwin, C. Manser, C. C. Miller, T. Hortobagyi, M. Dragunow, B. Rogelj, and C. E. Shaw. ALS mutant FUS disrupts nuclear localization and sequesters wild-type FUS within cytoplasmic stress granules. *Hum. Mol. Genet.*, 22(13):2676–2688, Jul 2013. 19

K. R. Veeramah, L. Johnstone, T. M. Karafet, D. Wolf, R. Sprissler, J. Salogiannis, A. Barth-Maron, M. E. Greenberg, T. Stuhlmann, S. Weinert, T. J. Jentsch, M. Pazzi, L. L. Restifo, D. Talwar, R. P. Erickson, and M. F. Hammer. Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia*, 54(7):1270–1281, Jul 2013. 214

J. Vengoechea, M. P. David, S. R. Yaghi, L. Carpenter, and S. A. Rudnicki. Clinical variability and female penetrance in X-linked familial FTD/ALS caused by a P506S mutation in UBQLN2. *Amyotroph Lateral Scler Frontotemporal Degener*, 14(7-8): 615–619, Dec 2013. 23

D. S. Verbeek, J. H. Schelhaas, E. F. Ippel, F. A. Beemer, P. L. Pearson, and R. J. Sinke. Identification of a novel SCA locus (SCA19) in a Dutch autosomal dominant cerebellar ataxia family on chromosome region 1p21-q21. *Hum. Genet.*, 111(4-5): 388–393, Oct 2002. 32

M. D. Vigeland. *paramlink: Parametric Linkage and Other Pedigree Analysis in R*, 2018. URL https://CRAN.R-project.org/package=paramlink. 65

D. Vilchez, I. Saez, and A. Dillin. The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nat Commun*, 5:5659, Dec 2014. 9, 10

K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.*, 55(4):641–658, Apr 2009. 33

J. Vogt, J. Kohlhase, S. Morlot, L. Kluwe, V. F. Mautner, D. N. Cooper, and H. Kehrer-Sawatzki. Monozygotic twins discordant for neurofibromatosis type 1 due to a postzygotic NF1 gene mutation. *Hum. Mutat.*, 32(6):E2134–2147, Jun 2011. 34, 222

I. F. Wang, L. S. Wu, H. Y. Chang, and C. K. Shen. TDP-43, the signature protein of FTLD-U, is a neuronal activity-responsive factor. *J. Neurochem.*, 105(3):797–806, May 2008. 16

K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164, Sep 2010. 48, 64, 80, 289

G. R. Warnes, B. Bolker, G. Gorjanc, G. Grothendieck, A. Korosec, T. Lumley, D. MacQueen, A. Magnusson, J. Rogers, and others. *gdata: Various R Programming Tools for Data Manipulation*, 2017. URL https://CRAN.R-project.org/package=gdata. 65

S. T. Warraich, S. Yang, G. A. Nicholson, and I. P. Blair. TDP-43: a DNA and RNA binding protein with roles in neurodegenerative diseases. *Int. J. Biochem. Cell Biol.*, 42(10):1606–1609, Oct 2010. 16

B. Weckselblatt and M. K. Rudd. Human Structural Variation: Mechanisms of Chromosome Rearrangements. *Trends Genet.*, 31(10):587–599, Oct 2015. 272, 282

F. B. Weihmuller, J. Ulas, L. Nguyen, C. W. Cotman, and J. F. Marshall. Elevated NMDA receptors in parkinsonian striatum. *Neuroreport*, 3(11):977–980, Nov 1992. 215

N. I. Weisenfeld, S. Yin, T. Sharpe, B. Lau, R. Hegarty, L. Holmes, B. Sogoloff, D. Tabbaa, L. Williams, C. Russ, C. Nusbaum, E. S. Lander, I. MacCallum, and D. B. Jaffe. Comprehensive variation discovery in single human genomes. *Nat. Genet.*, 46(12): 1350–1355, Dec 2014. 268, 269, 290, 291, 293

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL http://ggplot2.org. 65

H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2018. URL https://CRAN.R-project.org/package=stringr. 65

H. Wickham, J. Hester, and R. Francois. *readr: Read Rectangular Text Data*, 2017. URL https://CRAN.R-project.org/package=readr. 65

H. Wickham, R. Franois, L. Henry, and K. Mller. *dplyr: A Grammar of Data Manipulation*, 2018. URL https://CRAN.R-project.org/package=dplyr. 65

J. E. Wigginton and G. R. Abecasis. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics*, 21(16):3445–3447, Aug 2005. 180

J. R. Williams, E. Trias, P. R. Beilby, N. I. Lopez, E. M. Labut, C. S. Bradford, B. R. Roberts, E. J. McAllum, P. J. Crouch, T. W. Rhoads, C. Pereira, M. Son, J. L. Elliott, M. C. Franco, A. G. Estevez, L. Barbeito, and J. S. Beckman. Copper delivery to the CNS by CuATSM effectively treats motor neuron disease in SOD(G93A) mice co-expressing the Copper-Chaperone-for-SOD. *Neurobiol. Dis.*, 89:1–9, May 2016a. 144

K. L. Williams. *Encyclopedia of Bioinformatics and Computational Biology*, volume 2, chapter Gene Mapping. Elsevier, 2018. 28

K. L. Williams, J. A. Solski, G. A. Nicholson, and I. P. Blair. Mutation analysis of VCP in familial and sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging*, 33(7): 15–16, Jul 2012a. 169

K. L. Williams, S. T. Warraich, S. Yang, J. A. Solski, R. Fernando, G. A. Rouleau, G. A. Nicholson, and I. P. Blair. UBQLN2/ubiquilin 2 mutation and pathology in familial amyotrophic lateral sclerosis. *Neurobiol. Aging*, 33(10):3–10, Oct 2012b. 10, 23, 24, 92, 205

K. L. Williams, J. A. Fifita, S. Vucic, J. C. Durnall, M. C. Kiernan, I. P. Blair, and G. A. Nicholson. Pathophysiological insights into ALS with C9ORF72 expansions. *J. Neurol. Neurosurg. Psychiatr.*, 84(8):931–935, Aug 2013. 21, 148, 278, 279

K. L. Williams, E. P. McCann, J. A. Fifita, K. Zhang, E. L. Duncan, P. J. Leo, M. Marshall, D. B. Rowe, G. A. Nicholson, and I. P. Blair. Novel TBK1 truncating mutation in a familial amyotrophic lateral sclerosis patient of Chinese origin. *Neurobiol. Aging*, 36(12):1–5, Dec 2015. 10, 205, 282

K. L. Williams, S. Topp, S. Yang, B. Smith, J. A. Fifita, S. T. Warraich, K. Y. Zhang, N. Farrawell, C. Vance, X. Hu, A. Chesi, C. S. Leblond, A. Lee, S. L. Rayner, V. Sundaramoorthy, C. Dobson-Stone, M. P. Molloy, M. van Blitterswijk, D. W. Dickson, R. C. Petersen, N. R. Graff-Radford, B. F. Boeve, M. E. Murray, C. Pottier, E. Don, C. Winnick, E. P. McCann, A. Hogan, H. Daoud, A. Levert, P. A. Dion, J. Mitsui, H. Ishiura, Y. Takahashi, J. Goto, J. Kost, C. Gellera, A. S. Gkazi, J. Miller, J. Stockton, W. S. Brooks, K. Boundy, M. Polak, J. L. Munoz-Blanco, J. Esteban-Perez, A. Rabano, O. Hardiman, K. E. Morrison, N. Ticozzi, V. Silani, J. de Belleroche, J. D. Glass, J. B. Kwok, G. J. Guillemin, R. S. Chung, S. Tsuji, R. H. Brown, A. Garcia-Redondo, R. Rademakers, J. E. Landers, A. D. Gitler, G. A. Rouleau, N. J. Cole, J. J. Yerbury, J. D. Atkin, C. E. Shaw, G. A. Nicholson, and I. P. Blair. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nat Commun*, 7:11253, Apr 2016b. 10, 12, 25, 35, 38, 148, 167, 169, 205, 272, 278, 282

C. H. Wong, S. Topp, A. S. Gkazi, C. Troakes, J. W. Miller, M. de Majo, J. Kirby, P. J. Shaw, K. E. Morrison, J. de Belleroche, C. A. Vance, A. Al-Chalabi, S. Al-Sarraj, C. E. Shaw, and B. N. Smith. The CHCHD10 P34S variant is not associated with ALS in a UK cohort of familial and sporadic patients. *Neurobiol. Aging*, 36(10): 17–18, Oct 2015. 106, 107

P. C. Wong, C. A. Pardo, D. R. Borchelt, M. K. Lee, N. G. Copeland, N. A. Jenkins, S. S. Sisodia, D. W. Cleveland, and D. L. Price. An adverse property of a familial ALS-linked SOD1 mutation causes motor neuron disease characterized by vacuolar degeneration of mitochondria. *Neuron*, 14(6):1105–1116, Jun 1995. 14

C. H. Wu, C. Fallini, N. Ticozzi, P. J. Keagle, P. C. Sapp, K. Piotrowska, P. Lowe, M. Koppers, D. McKenna-Yasek, D. M. Baron, J. E. Kost, P. Gonzalez-Perez, A. D. Fox, J. Adams, F. Taroni, C. Tiloca, A. L. Leclerc, S. C. Chafe, D. Mangroo, M. J. Moore, J. A. Zitzewitz, Z. S. Xu, L. H. van den Berg, J. D. Glass, G. Siciliano, E. T.

Cirulli, D. B. Goldstein, F. Salachas, V. Meininger, W. Rossoll, A. Ratti, C. Gellera, D. A. Bosco, G. J. Bassell, V. Silani, V. E. Drory, R. H. Brown, and J. E. Landers. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature*, 488(7412):499–503, Aug 2012a. 12, 25, 34

L. S. Wu, W. C. Cheng, and C. K. Shen. Targeted depletion of TDP-43 expression in the spinal cord motor neurons leads to the development of amyotrophic lateral sclerosis-like phenotypes in mice. *J. Biol. Chem.*, 287(33):27335–27344, Aug 2012b. 16

Q. Wu, M. Liu, C. Huang, X. Liu, B. Huang, N. Li, H. Zhou, and X. G. Xia. Pathogenic Ubqln2 gains toxic properties to induce neuron death. *Acta Neuropathol.*, 129(3): 417–428, Mar 2015. 24

Z. Xi, L. Zinman, D. Moreno, J. Schymick, Y. Liang, C. Sato, Y. Zheng, M. Ghani, S. Dib, J. Keith, J. Robertson, and E. Rogaeva. Hypermethylation of the CpG island near the G4C2 repeat in ALS with a C9orf72 expansion. *Am. J. Hum. Genet.*, 92 (6):981–989, Jun 2013. 284

Z. Xi, M. Zhang, A. C. Bruni, R. G. Maletta, R. Colao, P. Fratta, J. M. Polke, M. G. Sweeney, E. Mudanohwo, B. Nacmias, S. Sorbi, M. C. Tartaglia, I. Rainero, E. Rubino, L. Pinessi, D. Galimberti, E. I. Surace, P. McGoldrick, P. McKeever, D. Moreno, C. Sato, Y. Liang, J. Keith, L. Zinman, J. Robertson, and E. Rogaeva. The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol.*, 129(5):715–727, May 2015. 284

Z. Xu, M. Poidevin, X. Li, Y. Li, L. Shu, D. L. Nelson, H. Li, C. M. Hales, M. Gearing, T. S. Wingo, and P. Jin. Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. *Proc. Natl. Acad. Sci. U.S.A.*, 110(19):7778–7783, May 2013. 22

Y. Xue, Y. Chen, Q. Ayub, N. Huang, E. V. Ball, M. Mort, A. D. Phillips, K. Shaw, P. D. Stenson, D. N. Cooper, and C. Tyler-Smith. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.*, 91(6):1022–1032, Dec 2012. 296

M. Yandell, C. Huff, H. Hu, M. Singleton, B. Moore, J. Xing, L. B. Jorde, and M. G. Reese. A probabilistic disease-gene finder for personal genomes. *Genome Res.*, 21 (9):1529–1542, Sep 2011. 38

S. Yang, S. T. Warraich, G. A. Nicholson, and I. P. Blair. Fused in sarcoma/translocated in liposarcoma: a multifunctional DNA/RNA binding protein. *Int. J. Biochem. Cell Biol.*, 42(9):1408–1411, Sep 2010. 18

Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, 369(16):1502–1511, Oct 2013. 32

Y. Yang, D. M. Muzny, F. Xia, Z. Niu, R. Person, Y. Ding, P. Ward, A. Braxton, M. Wang, C. Buhay, N. Veeraraghavan, A. Hawes, T. Chiang, M. Leduc, J. Beuten, J. Zhang, W. He, J. Scull, A. Willis, M. Landsverk, W. J. Craigen, M. R. Bekheirnia, A. Stray-Pedersen, P. Liu, S. Wen, W. Alcaraz, H. Cui, M. Walkiewicz, J. Reid, M. Bainbridge, A. Patel, E. Boerwinkle, A. L. Beaudet, J. R. Lupski, S. E. Plon, R. A. Gibbs, and C. M. Eng. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879, Nov 2014. 32

K. Ye, M. Beekman, E. W. Lameijer, Y. Zhang, M. H. Moed, E. B. van den Akker, J. Deelen, J. J. Houwing-Duistermaat, D. Kremer, S. Y. Anvar, J. F. Laros, D. Jones, K. Raine, B. Blackburne, S. Potluri, Q. Long, V. Guryev, R. van der Breggen, R. G. Westendorp, P. A. 't Hoen, J. den Dunnen, G. J. van Ommen, G. Willemsen, S. J. Pitts, D. R. Cox, Z. Ning, D. I. Boomsma, and P. E. Slagboom. Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet*, 16(6):1026–1032, Dec 2013. 267, 268

P. E. Young, S. Kum Jew, M. E. Buckland, R. Pamphlett, and C. M. Suter. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. *PLoS ONE*, 12(8):e0182638, 2017. 266, 272, 284

X. Yu and S. Sun. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, 14:274, Sep 2013. 288, 291, 292, 293

Z. Yuan, B. Jiao, L. Hou, T. Xiao, X. Liu, J. Wang, J. Xu, L. Zhou, X. Yan, B. Tang, and L. Shen. Mutation analysis of the TIA1 gene in Chinese patients with amyotrophic lateral sclerosis and frontotemporal dementia. *Neurobiol. Aging*, 64:9–160, Apr 2018. 161, 278

F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*, 10:451–481, 2009. 305

K. Zhang, Q. Liu, D. Shen, H. Tai, H. Fu, S. Liu, Z. Wang, J. Shi, Q. Ding, X. Li, M. Liu, L. Cui, and X. Zhang. Genetic analysis of TIA1 gene in Chinese patients with amyotrophic lateral sclerosis. *Neurobiol. Aging*, 67:9–201, Jul 2018. 161, 278

M. Zhang, Z. Xi, M. Ghani, P. Jia, M. Pal, K. Werynska, D. Moreno, C. Sato, Y. Liang, J. Robertson, A. Petronis, L. Zinman, and E. Rogaeva. Genetic and epigenetic study of ALS-discordant identical twins with double mutations in SOD1 and ARHGEF28. *J. Neurol. Neurosurg. Psychiatry*, 87(11):1268–1270, 11 2016. 268, 272

X. Zhang. Exome sequencing greatly expedites the progressive research of Mendelian diseases. *Front Med*, 8(1):42–57, Mar 2014. 45, 52

Q. Zheng and E. A. Grice. AlignerBoost: A Generalized Software Toolkit for Boosting Next-Gen Sequencing Mapping Accuracy Using a Bayesian-Based Mapping Quality Framework. *PLoS Comput. Biol.*, 12(10):e1005096, Oct 2016. 291, 293

Q. Zhou, Y. Chen, Q. Wei, B. Cao, Y. Wu, B. Zhao, R. Ou, J. Yang, X. Chen, S. Hadano, and H. F. Shang. Mutation Screening of the CHCHD10 Gene in Chinese Patients with Amyotrophic Lateral Sclerosis. *Mol. Neurobiol.*, 54(5):3189–3194, Jul 2017. 106

Y. Zhu, C. Tazearslan, and Y. Suh. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Exp. Biol. Med. (Maywood)*, 242(13):1325–1334, 07 2017. 298

H. Zinszner, J. Sok, D. Immanuel, Y. Yin, and D. Ron. TLS (FUS) binds RNA in vivo and engages in nucleo-cytoplasmic shuttling. *J. Cell. Sci.*, 110 (Pt 15):1741–1750, Aug 1997. 18

J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.*, 32(3):246–251, Mar 2014. 226, 269, 290, 291

Z. Y. Zou, L. Y. Cui, Q. Sun, X. G. Li, M. S. Liu, Y. Xu, Y. Zhou, and X. Z. Yang. De novo FUS gene mutations are associated with juvenile-onset sporadic amyotrophic lateral sclerosis in China. *Neurobiol. Aging*, 34(4):1–8, Apr 2013. 18

Z. Y. Zou, Z. R. Zhou, C. H. Che, C. Y. Liu, R. L. He, and H. P. Huang. Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry*, 88(7):540–549, 07 2017. 279, 280

T. Zu, B. Gibbens, N. S. Doty, M. Gomes-Pereira, A. Huguet, M. D. Stone, J. Margolis, M. Peterson, T. W. Markowski, M. A. Ingram, Z. Nan, C. Forster, W. C. Low, B. Schoser, N. V. Somia, H. B. Clark, S. Schmechel, P. B. Bitterman, G. Gourdon, M. S. Swanson, M. Moseley, and L. P. Ranum. Non-ATG-initiated translation directed by microsatellite expansions. *Proc. Natl. Acad. Sci. U.S.A.*, 108(1):260–265, Jan 2011. 22

T. Zu, Y. Liu, M. Banez-Coronel, T. Reid, O. Pletnikova, J. Lewis, T. M. Miller, M. B. Harms, A. E. Falchook, S. H. Subramony, L. W. Ostrow, J. D. Rothstein, J. C. Troncoso, and L. P. Ranum. RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):E4968–4977, Dec 2013. 9, 22

P. J. Zwijnenburg, H. Meijers-Heijboer, and D. I. Boomsma. Identical but not the same: the value of discordant monozygotic twins in genetic research. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 153B(6):1134–1149, Sep 2010. 222